Generierung einer dynamischen Keimbahn-V-Gen-Datenbank und *in-silico*-Charakterisierung des Immunglobulin-Schwerekettenlocus des Mausstammes 129/Sv

Vom Fachbereich für Biowissenschaften und Psychologie
der Technischen Universität Carolo-Wilhelmina
zu Braunschweig
zur Erlangung des Grades einer
Doktorin der Naturwissenschaften
(Dr.rer.nat.)
genehmigte
D i s s e r t a t i o n

von Ida Retter
aus Mutlangen

1. Referent: Prof. Dr. Jürgen Wehland

2. Referent: Prof. Dr. Stefan Dübel

eingereicht am: 26.09.2005

mündliche Prüfung (Disputation) am: 08.02.2006

Vorveröffentlichungen der Dissertation

Teilergebnisse aus dieser Arbeit wurden mit Genehmigung der Gemeinsamen Naturwissenschaftlichen Fakultät, vertreten durch den Mentor der Arbeit, in folgenden Beiträgen vorab veröffentlicht:

Publikationen

Retter, I., Althaus, H.H., Münch, R., Müller, W.: VBASE2, an integrative V gene database.
Nucleic Acids Res. 2005 Jan. 1; 33(Database issue): D671-4.

Tagungsbeiträge

Retter, I., Blöcker, H., Hühne, R., Scharfe, M., Riblet, R. und Müller, W.: Characterisation of the immunoglobulin heavy chain locus of the mouse. Poster.
Jahrestagung der Gesellschaft für Immunologie 2003, Berlin, Deutschland.

Retter, I., Hühne, R. und Müller, W.: Automatic generation of a germ-line encoded immunoglobulin heavy chain variable gene database of the mouse from the EMBL nucleotide database. Poster.
Deutsche Bioinformatikkonferenz GCB 2003, München, Deutschland.

Retter, I. und Müller, W.: Providing an automatically derived high quality immunoglobulin Vgene sequence database. Poster.
International Conference on Research in Computational Molecular Biology RECOMB 2004,San Diego, USA.

Müller, W. und Retter, I.: VBASE2, an integrative V gene database. Poster.
HUGO's 10th Human Genome Meeting HGM 2005, Kyoto, Japan.

Retter, I., Chevillard, C., Ludewig, M., Mauhar, A., Nordsiek, G., Scharfe, M., Blöcker, H., Riblet, R. und Müller, W.: Elucidating the immunoglobulin heavy chain locus of the 129/Sv mouse strain. Poster.
Genetics of Infection. Jahrestagung der Deutschen Gesellschaft für Genetik 2005, Braunschweig, Deutschland.

Retter, I., Müller, W., O'Donovan, C., Martin, M.J. und Apweiler, R.: Selecting immunoglobulin proteins for the UniProt Knowledgebase. Poster.
Deutsche Bioinformatikkonferenz GCB 2005, Hamburg, Deutschland.

Retter, I., Chevillard, C., Ludewig, M., Mauhar, A., Nordsiek, G., Scharfe, M., Blöcker, H., Riblet, R. und Müller, W. : Characterisation of 1.5 Mb of the immunoglobulin heavy chain locus of the 129/Sv mouse. Poster.
International Mouse Genome Conference 2005, Straßburg, Frankreich.

Inhaltsverzeichnis

Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit der *in-silico*-Analyse von antikörperkodierenden DNA- Sequenzen. Zum einen wurde ein automatischer V-Gen-Analyse-Prozess entwickelt, mit dem die Datenbank VBASE2 erzeugt wird. Zum anderen wurde mit Unterstützung der automatischen V-Gen-Analyse und weiteren bioinformatischen Methoden die genomische Sequenz des Immunglobulin-Schwerekettenlocus (IgH-Locus) des Mausstammes 129/Sv untersucht und annotiert.

Der Prozess zur automatischen V-Gen-Analyse führt eine systematische Sortierung und Klassifizierung der V-Gen-Sequenzen der EMBL-Bank durch. Die durch den Prozess generierte Datenbank VBASE2 ist eine integrative Keimbahn-V-Gen-Datenbank, deren Einträge Verweise auf die EMBL-Bank, andere immunologische Datenbanken und auf Ensembl enthalten. Über den Ensembl DAS Server können VBASE2-V-Gene im Ensembl Genombrowser dargestellt werden. Die V-Gen-Einträge von VBASE2 sind in Klassen unterschiedlicher Qualität eingeteilt, die die Zuverlässigkeit der jeweiligen Sequenz widerspiegeln. Der aktuelle Datensatz von VBASE2 beinhaltet 1129 Keimbahn-V-Gene, -Pseudogene, V-Gen-Relikte und Orphans der Immunglobulinloci von Mensch und Maus. Auf der Grundlage des VBASE2-Generierungsprozesses wurde ein Modul für die Proteindatenbank UniProtKB/TrEMBL entwickelt, das den Eingang von Immunglobulinsequenzen in diese automatisch erzeugte Datenbank kontrolliert. Weiterhin wurde der Prozess für die Annotation der V-Segmente des IgH-Locus des Mausstammes 129/Sv genutzt. Der in der vorliegenden Arbeit untersuchte Teil des IgH-Locus umfasst mit 1,4 Mb die konstante Region, D-Region, J-Region und etwa ein Drittel der variablen Region. Durch die Annotation dieser Sequenz konnten neue V- und D-Segmente beschrieben werden. Vier V-Gen-Relikte und ein funktionelles V-Gen der V_h7183- und der V_hQ52-Familie liegen jeweils als exakte Duplikationen vor. Sequenzvergleiche der D-Region von 129/Sv zeigen deutliche Unterschiede sowohl zum Igh^a- (Stamm BALB/c) als auch zum Igh^b- (Stamm C57BL/6) Haplotyp. Eine Untersuchung der repetitiven Elemente im IgH-Locus von 129/Sv ergab unter anderem einen ungewöhnlich hohen Anteil an LINE1-Elementen. Die hier beschriebene Charakterisierung des IgH-Locus der 129/Sv-Maus setzt die am Mausstamm C57BL/6 begonnene Arbeit fort und trägt zur Aufklärung dieses komplexen Locus im Modellorganismus Maus bei.

1. Einleitung

1.1 Einführung in die Struktur und Funktionsweise der Immunglobulinloci

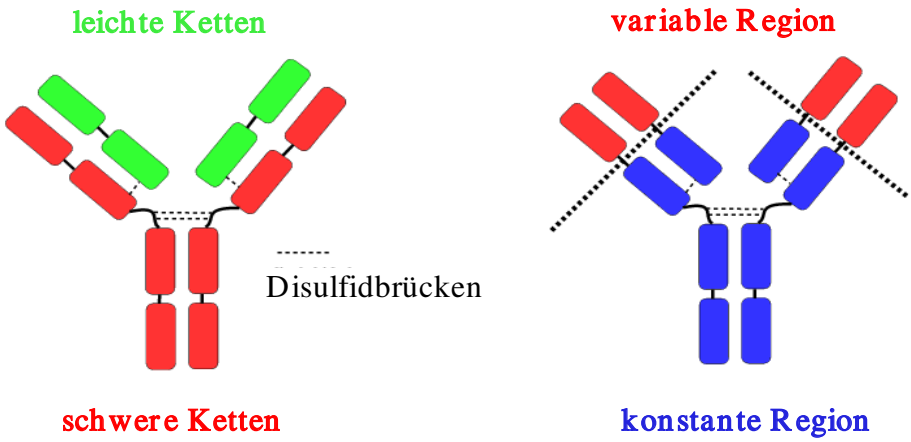
Das Immunsystem der Vertebraten verteidigt das Individuum gegenüber Angriffen pathogener Organismen. Wird die Existenz von fremden Zellen oder Molekülen im Körper auf molekularer Ebene erkannt, löst dies eine Immunantwort aus, welche die Vernichtung der Fremdkörper zum Ziel hat. Dabei verfolgt das Immunsystem zwei verschiedene Abwehr-Strategien: die eine Strategie beruht auf der Erkennung fremder Strukturen durch Mechanismen, die dem Organismus angeboren sind. Diese angeborene Immunantwort erfolgt sofort nach einer Infektion, ist jedoch wenig spezifisch und für die vollständige Beseitigung der Pathogene oft nicht ausreichend. Die andere Strategie wird bestimmt durch Moleküle, die in einem länger dauernden Selektionsverfahren so verändert werden, dass sie körperfremde Bestandteile hochspezifisch erkennen und eine gezielte Vernichtungsreaktion auslösen. Diese erworbene oder adaptive Immunantwort beruht auf einem Prozess, in dem das Immunsystem im Verlauf der Individualentwicklung dazu ausgebildet wird, immer mehr Pathogene zu erkennen und sich über viele Jahre hinweg an sie zu erinnern.

Auf zellulärer Ebene wird die adaptive Immunantwort durch die Lymphocyten, eine Form der weißen Blutkörperchen, vermittelt. Man unterscheidet zwischen B- und T-Lymphozyten: die B-Lymphozyten werden im Knochenmark gebildet und produzieren Immunglobuline. Diese Proteine sitzen als B-Zell-Rezeptoren auf der Zelloberfläche und können von spezialisierten B-Zellen in großer Menge sekretiert werden. Immunglobuline dienen der spezifischen Erkennung von körperfremden Molekülen und werden auch Antikörper genannt. Ein Molekül,

das von einem Antikörper erkannt wird, wird als Antigen bezeichnet. T-Lymphozyten reifen im Thymus heran und tragen auf ihrer Oberfläche T-Zell-Rezeptoren, die den B-Zell-Rezeptoren strukturell ähnlich sind und ebenfalls der spezifischen Antigen-Erkennung dienen. Im Gegensatz zu den Immunglobulinen gibt es von T-Zell-Rezeptoren jedoch keine sekretierte Form.

B- und T-Zellrezeptoren bestehen aus konstanten und variablen Bereichen, die jeweils durch eigene Proteindomänen repräsentiert werden. Die variablen Regionen sind für die Antigenbindung zuständig und unterscheiden sich bei den Rezeptoren verschiedener Lymphozyten. Eine Antigenbindungsstelle wird jeweils aus zwei verschiedenen Peptidketten gebildet. T-Zell-Rezeptoren haben nur eine Antigenbindungsstelle und sind Heterodimere. B-Zell-Rezeptoren haben zwei Antigenbindungsstellen und sind Tetramere (Abbildung 1.1). Sie bestehen aus zwei identischen schweren und zwei identischen leichten Ketten, die durch Disulfidbrücken kovalent miteinander verbunden sind. Die Antigenbindungsstellen werden jeweils durch schwere und leichte Kette gebildet. Die konstante Region ist für die Interaktion mit anderen Komponenten des Immunsystems verantwortlich. Dabei werden durch verschiedene Isotypen von schweren Ketten unterschiedliche Antikörper-Klassen mit diversen Effektorfunktionen gebildet. Bei der Maus gibt es die Klassen IgM, IgD, IgG_3, IgG_1, IgG_{2b}, IgG_{2a}, IgE und IgA; die konstanten Regionen der dazugehörigen schweren Ketten werden durch die entsprechenden griechischen Buchstaben bezeichnet.

Abbildung 1.1: Grundstruktur eines Antikörpers



Quelle: http://www-immuno.path.cam.ac.uk/ (angepasst)

Die Notwendigkeit, jedes beliebige Pathogen durch eine spezifische Bindung zu erkennen, erfordert eine enorme strukturelle Diversität der variablen Regionen. Diese wird zum einen durch eine große Anzahl multipler Gensegmente und Allele in der Keimbahn-DNA erreicht. Zum anderen sind für die Bildung von Antikörpern somatische Veränderungen der DNA nötig [Review: Tonegawa, 1983]. Aus diesem Grund haben die Immunglobulinloci, ebenso wie die ähnlich gebauten T-Zell-Rezeptor-Loci, eine ganz besondere Struktur, bei der die kodierende Sequenz erst in der frühen B- und T-Zellentwicklung gebildet wird [Buch: Honjo und Alt (Ed.), 1995].

Bei der Maus gibt es wie beim Menschen einen Schwerekettenlocus (IgH-Locus) und zwei Leichtekettenloci, den Kappa- und den Lambda-Locus, an denen zwei verschiedene leichte Ketten gebildet werden. Alle drei Loci enthalten multiple V(variable)-Segmente und J(joining)-Segmente. Der Schwerekettenlocus enthält zusätzlich D(diversity)-Segmente, die zwischen den V- und den J-Segmenten liegen (Abbildung 1.2) [Early et al., 1980]. V-Segmente sind etwa 300 Nukleotide lang, kodieren für den größten Teil der variablen Region und werden auch als V-Gene bezeichnet. Sie bestehen aus zwei Exonen, wobei das erste Exon und der Beginn des zweiten Exons für ein Signalpeptid zur Aufnahme in

das Endoplasmatische Retikulum kodieren [Milstein et al., 1972; Early et al., 1980]. D- und J-Segmente sind mit etwa zehn bis fünfzig Nukleotiden deutlich kürzer als V-Segmente [Newell et al., 1980; Ye, 2004].

Abbildung 1.2: VDJ-Rekombination am IgH-Locus

Schematische Darstellung des murinen IgH-Locus. Kodierende Segmente werden durch Vierecke, Transkription durch Pfeile symbolisiert. Die C-Region ist von zwei durch Kreise dargestellten Enhancer-Elementen, Eiμ und 3'Eμ, umgeben.

A: Die Keimbahn-Konfiguration. 23-RSS sind als schwarze, 12-RSS als weiße Dreiecke dargestellt.

B: Ablauf der VDJ-Rekombination. Promotoren sind als Elipsen dargestellt. Zunächst wird ein D- und ein J-Segment miteinander verbunden, so dass ein Dμ-Protein gebildet werden kann. Durch anschließende Verbindung des DJ-Rearrangements mit einem V-Segment kann ein vollständiges kodierendes Transkript gebildet werden.

A

Quelle: Hesslein und Schatz, 2001

Durch eine somatische DNA-Rekombination wird jeweils ein V-, ein D- und ein J-Segment (beziehungsweise ein V- und ein J-Segment am Kappa- und Lambda-Locus) miteinander verbunden, wobei die dazwischen befindliche DNA

ausgeschnitten wird (Abbildung 1.2) [Reviews: Schatz, 2004; Bassing et al., 2002; Gellert, 2002; Hesslein und Schatz, 2001]. Diese V(D)J-Rearrangements sind die kodierenden Sequenzen für die variablen Regionen der schweren und leichten Kette des Antikörpers. Die V(D)J-Rekombination findet im Knochenmark und in der fötalen Leber statt. Der Rekombinationsprozess wird durch Signalsequenzen (RSS-Elemente) vermittelt, die sich am 3'-Ende der V-Segmente, am 5'-Ende der J-Segmente und beiderseits der D-Segmente befinden [Early et al., 1980]. Ein RSS-Element besteht aus einem konserviertem Heptamer-Motiv, einem wenig konservierten Abstandhalter (Spacer) definierter Länge und einem Nonamer-Motiv. Der Spacer ist entweder 12 +/-1 bp oder 23 +/-1 bp lang. Für die Rekombination muss ein RSS-Element mit einem 12-bp-Spacer (benannt als 12-RSS) und ein RSS-Element mit einem 23-bp-Spacer (benannt als 23-RSS) zusammenkommen. Diese 12/23-Regel erlaubt die geordnete Rekombination von drei Sorten von Segmenten, da am IgH-Locus alle V-Gene ein 23-RSS tragen, die D-Segmente beiderseits 12-RSS und die J-Segmente widerum ein 23-RSS. Die RSS-Elemente dienen einem speziellen DNA-Rekombinase-Komplex als Erkennungssequenzen [Review: Swanson, 2004]. Hauptbestandteile der Rekombinase sind die Enzyme RAG1 und RAG2, die den Doppelstrangbruch induzieren, eine Haarnadelstruktur an den DNA-Enden bilden und die Enden bis zur Religation gebunden halten [Schatz et al., 1989; Oettinger et al., 1990; Reviews: Brandt und Roth, 2004; Roth, 2003]. Versetzes Schneiden der Haarnadelstruktur und das Enzym terminale desoxyNucleotidyl-Transferase sorgen während der Rekombination für zusätzliche Diversität in Form von P(palindromic)- und N(non-template)-Nukleotiden im Bereich der Schnittstelle [Review: Lewis, 1994]. Die Religation der DNA-Doppelstränge erfolgt durch das “nonhomologous DNA end joining” (NHEJ). Dieser hochkonservierte Mechanismus ist auch bei Evertebraten für die Reparatur von Doppelstrangbrüchen zuständig, die durch ionisierende

Strahlung oder oxidative Prozesse entstehen können [Review: Lieber et al., 2004].

Voraussetzung für die V(D)J-Rekombination ist die Transkription des zu rekombinierenden Bereichs. Promotorregionen liegen am murinen IgH-Locus auf der 5'-Seite der V-Segmente [Reviews: Henderson und Calame, 1998; Schebesta et al., 2002; Johnson et al., 2005], der D-Segmente und der Isotypen der konstanten Regionen (siehe Abbildungen 1.2 und 1.3). Weiterhin gibt es zwei Enhancer-Elemente, die die konstante Region einrahmen: der "intronic enhancer" (Eiμ) liegt 5' des ersten Cμ-Exons im J_H Intron, der "3' enhancer" (E3') liegt auf der 3'-Seite der konstanten Region. Vor Beginn der VDJ-Rekombination ist der Promotor des DQ52 -D-Segments (PDQ52) aktiv, das der J-Region benachbart liegt. PDQ52 induziert die Bildung der sterilen Transkripte μ^0 und Iμ (siehe Abbildung 1.2). Im anschließenden ersten Rekombinationsschritt wird ein D-Segment mit einem J-Segment verbunden. Nun erfolgt die Bildung eines Dμ-Transkripts vom Promotor des rearrangierten D-Segments aus. In Abhängigkeit vom Leseraster des D-Segments kann daraus ein verkürztes μ-Protein gebildet werden, wenn das Leseraster des D-Segments mit dem Raster der Translationsstartstelle zu Beginn des Transkripts übereinstimmt. Die Bildung dieses sogenannten Dμ-Proteins führt zur Inhibierung weiterer Rearrangements in der Zelle, wodurch das Leseraster des D-Segments selektioniert wird [Reth und Alt, 1984; Gu et al., 1991; Ehlich et al., 1994]. Bei korrektem D-Segment-Leseraster findet anschließend das V-DJ-Rearrangement statt, wobei nun sterile Transkripte von den Promotoren der V-Segmente aus gebildet werden. Neben V-Gen-Transkripten in kodierender Richtung werden in nichtkodierender Richtung Transkripte über mehrere V-Gene hinweg gebildet [Bolland et al., 2004]. Das erfolgreiche Rearrangement der schweren Kette bringt den Promotor des V-Segments in räumliche Nähe

zum Eiμ-Enhancer. Dies führt zur verstärkten Transkription und zur Bildung des vollständigen VDJμ-Proteins, das in Verbindung mit einer Ersatz-Leichten-Kette auf der Zelloberfläche als preB-Zell-Rezeptor exprimiert wird. Mit der Expression des preB-Zell-Rezeptors auf der Zelloberfläche ist das sogenannte Pro-B-Zell-Stadium und das Rearrangement der Schwerekettenloci beendet. Im anschließenden Pre-B-Zell-Stadium rearrangiert einer der Leichtekettenloci, dessen erfolgreiches Rearrangement schließlich zur B-Zell-Rezeptor-Expression und damit zum Übergang in den Zustand der unreifen B-Zelle führt.

Der B-Zell-Rezeptor wird zunächst auf Selbsttoleranz geprüft [Review: Nossal, 1992]. Die Bindung des B-Zell-Rezeptors an ein Antigen im Knochenmark induziert ein sekundäres Rearrangement der leichten Ketten; dieses so genannte “receptor editing” kann nach erfolglosem Rearrangement der beiden gewöhnlich zuerst rearrangierten Kappa-Allele auch zum Rearrangement der Lambda-Allele führen. Erst bei erfolglosem Rearrangement aller vier Leichte-Ketten-Allele bekommt die Zelle das Signal zur Apoptose, um autoreaktive Antikörper zu verhindern.

Das primäre Antikörper-Repertoire wird folglich unabhängig von körperfremden Antigenen gebildet [Review: Feeney, 2000; Alt et al., 1987]. Dabei verändert sich das Repertoire im Verlauf der Entwicklung: Während die in der fötalen Leber gebildeten B1-Zellen ein eingeschränktes Rezeptor-Repertoire aufweisen [Review: Hayakawa und Hardy, 2000], repräsentiert das Repertoire der im adulten Knochenmark gebildeten B2-Zellen das gesamte V-Gen-Spektrum. Zahlreiche Untersuchungen belegen, dass die Auswahl der Segmente auch in B2-Zellen nicht zufällig ist [Reviews: Feeney et al., 2004; Krangel, 2003; Nemazee und Hogquist, 2003]. Die Heptamer- und Nonamer-Konsensus-Sequenzen und die Spacer der RSS-Elemente beeinflussen die

Rekombinationshäufigkeit ebenso wie die angrenzenden kodierenden Sequenzen. Auch die Promotorregionen haben Einfluss auf die Bildung des primären Repertoires. In der späteren Entwicklung wird das Antikörper-Repertoire von B-Zellen verschiedener Subtypen und Kompartimente durch Liganden selektioniert [Gu et al., 1991; Review: Rajewsky, 1996].

Interessanterweise wird jeweils nur eines der beiden Allele exprimiert, im Falle der leichten Kette auch nur entweder ein Allel der Kappa-, oder ein Allel der Lambda-Kette. Dieses Phänomen wird als “allele Exklusion” bezeichnet und gewährleistet die Monospezifität aller Antigenrezeptoren einer Zelle [Reviews: Corcoran, 2005; Mostoslavsky et al., 2004; Bergman und Cedar, 2004; Goldmit und Bergman, 2004]. Funktionelle bispezifische Antikörper sind bisher nur in Ausnahmefällen bekannt [diskutiert in Melchers, 2004]. Dabei verläuft die DJ-Rekombination im frühen Pro-B-Zell-Stadium noch synchron auf beiden Allelen. Die anschließende V-DJ-Rekombination findet jedoch gewöhnlich nur dann auf dem zweiten Allel statt, wenn das erste Rearrangement eine Leserasterverschiebung trägt. Als Ursache für die asynchrone V-DJ-Rekombination und die anschließende asynchrone V-J-Rekombination der leichten Kette werden verschiedene Modelle diskutiert. Einigkeit besteht darüber, dass das Signal zur Beendigung der Rekombination im Sinne einer Feedback-Hemmung vom Rekombinationsprodukt, also vom Pre-B-Zell-Rezeptor beziehungsweise vom B-Zell-Rezeptor auf der Zelloberfläche ausgeht.

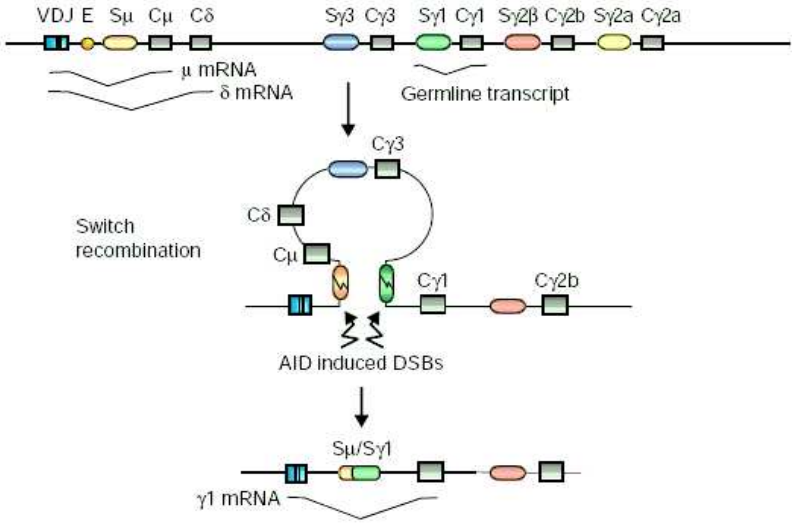
Der Ablauf der V(D)J-Rekombination wird also nicht nur zell- und entwicklungsspezifisch, sondern auch locus- und allelspezifisch exakt reguliert [Reviews: Oettinger, 2004; Schlissel, 2004]. Die Spezifität dieser Regulation ist nicht nur für die geordnete B- und T-Zellbildung von Bedeutung, sondern für den gesamten Organismus: Im Fall von fehlgeleiteten

Rekombinationsreaktionen kann es zu Chromosomentranslokationen und Onkogenaktivierung kommen, deren Auswirkungen leicht zum Tod des Individuums führen können [Reviews: Küppers und Dalla-Favera, 2001; Davila et al., 2001]. Die Regulation der V(D)J-Rekombination findet zum Teil durch Regulation der Transkription der RAG-Gene statt. Diese werden während der Rekombination exprimiert, nach erfolgreichem Rearrangement jedoch strikt inhibiert, wodurch sich die Zell- und Entwicklungspezifität des Prozesses erklären läßt. Für die Locus- und Allel-Spezifität sind jedoch epigenetische Mechanismen nötig, die auf Ebene der Chromatinstruktur die Öffnung oder Stilllegung der Immunglobulinloci ermöglichen [Reviews: Su und Tarakhovsky, 2005; Corcoran, 2005; Bergman und Cedar, 2004; Goldmit und Bergman, 2004]. Zur schrittweisen Öffnung der Loci für die Rekombination tragen die Relokalisation in das Zentrum des Zellkerns [Kosak et al., 2002], Acetylierung und Methylierung der Histone sowie die Demethylierung der DNA bei. Ob die Bildung bidirektionaler steriler Transkripte in der V-Region aktiv an der Öffnung des Locus beteiligt oder aber die Folge davon ist, ist noch unklar. Für die Rekombination der distalen V-Gene ist eine Kontraktion der Immunglobulinloci nötig [Fuxa et al., 2004; Roldan et al., 2005; Goldmit et al., 2005]. Offen ist jedoch, durch welche Mechanismen diese Vorgänge kontrolliert werden.

Ein weiteres somatisches DNA-Rekombinationsereignis erlaubt den Wechsel des Schwerenketten-Isotyps. Dabei wird ein VDJ-Rearrangement, das zunächst als IgM oder IgD exprimiert wird, mit der kodierenden Sequenz eines anderen Isotyps verbunden [Reviews: Chaudhuri und Alt, 2004; Stavnezer und Amemiya, 2004; Wang und Wabl, 2004; Kenter, 2003; Yu und Lieber, 2003]. Dieses Verfahren ermöglicht den Wechsel der Immunglobulinklasse ohne Veränderung der Antigenspezifität. Der Klassenwechsel verläuft jedoch nach einem grundsätzlich anderen Mechanismus als die V(D)J-Rekombination.

Abbildung 1.3: Klassenwechsel am murinen IgH-Locus

Oben auf dem Bild ist ein VDJ-Rearrangement und die Keimbahn-Konfiguration der konstanten Region mit Bildung der VDJμ- und VDJδ-mRNAs und eines sterilen Transkripts der g1-Region schematisch dargestellt. Im nächsten Schritt werden die S-Regionen von μ und g1 in räumliche Nähe gebracht. Nach AID-induzierten Doppelstrangbrüchen innerhalb der Sμ- und Sg1-Regionen kommt es zur Ligation der S-Regionen unter Deletion der dazwischen befindlichen DNA.



Quelle: Kenter, 2003

Die acht Isotypen der Maus werden durch acht Gruppen von Schwereketten-Exonen am 3'-Ende des IgH-Locus repräsentiert (Abbildung 1.3) [Shimizu et al., 1982]. Die einzelnen Domänen der schweren Kette werden dabei jeweils durch getrennte Exone kodiert. Angrenzend an die J-Region befinden sich die Exone der μ-Kette, gefolgt von den Exonen der δ-Kette. Die alternative Expression dieser beiden Isotypen wird durch alternative Polyadenylierung und mRNA-Prozessierung ermöglicht [Maki et al., 1981]. Für den Wechsel zu einer anderen Immunglobulin-Klasse rekombiniert die DNA im 5'-Bereich des ersten μ-Exons mit einem Bereich am 5'-Ende eines anderen Isotypen, wobei die dazwischen befindliche DNA ausgeschnitten wird [Honjo und Kataoka, 1978]. Die Sequenz im Rekombinationsbereich wird als S(switch)-Region bezeichnet und besteht aus sich wiederholenden Motiven unterschiedlicher Länge und Komplexität

[Review: Yu und Lieber, 2003]. 3' von der S-Region liegt ein Exon, das nicht für die schwere Kette kodiert, und ein Promotor. Dieses sogenannte I-Exon liegt am Beginn eines sterilen Transkripts, dessen Bildung Voraussetzung für den Klassenwechsel ist. Die Regulation des Klassenwechsels erfolgt also zumindest teilweise durch Transkriptionsfaktoren, die durch extrazelluläre Signale aktiviert werden und im Promotorbereich des I-Exons binden [Lee et al., 2001; Review: Lorenz et al., 1995]. I-Exon und S-Region liegen im 5'-Bereich aller Isotypen-Exone mit Ausnahme der δ-Region. Der biochemische Mechanismus des Klassenwechsels ist noch nicht vollständig aufgeklärt. Fest steht, dass das Enzym Aktivierungsinduzierte Cytidindeaminase (AID) eine wichtige Rolle bei der Bildung des Doppelstrangbruchs spielt.

Die AID ist ein Schlüsselenzym bei der somatischen Diversifikation von Immunglobulin-Genen, da es ebenfalls an der somatischen Hypermutation sowie an der Genkonversion von Rearrangements beteiligt ist [Reviews: Maizels, 2005; Honjo et al., 2004; Reynaud et al., 2003]. Diese drei Prozesse sind B-Zell-spezifische Diversifikationsmechanismen, und die Entdeckung der B-Zell-spezifisch exprimierten AID hat in den letzten Jahren sehr viele Untersuchungen zu den Gemeinsamkeiten dieser Mechanismen angeregt. Somatische Hypermutation tritt bei der sogenannten Affinitätsreifung in den Keimzentren der sekundären lymphatischen Organe auf, wenn die B-Zellen nach Bindung eines Antigens durch T-Zellen stimuliert werden [Review: Rajewsky, 1996]. Durch die Affinitätsreifung werden die Antikörper aktivierter B-Zellen durch Punktmutationen in der variablen Region diversifiziert, und Klone mit höherer Antigen-Affinität werden positiv selektioniert. Die AID ist dabei offensichtlich für die Induktion der Mutationen verantwortlich; der Mechanismus und insbesondere das Substrat der AID, DNA oder RNA, ist aber noch umstritten und wird derzeit intensiv untersucht [Reviews (Auswahl):

Neuberger et al., 2005; Petersen-Mahrt, 2005; Diaz und Lawrence, 2005; Franklin und Blanden, 2004]. Genkonversion ist ein Mechanismus, der unter anderem bei Vögeln und vielen Säugetieren neben der Hypermutation die Grundlage für die Erzeugung der Antikörperdiversität bildet [Review: Arakawa und Buerstedde, 2004].

Mit der zunehmenden Verfügbarkeit vollständiger Genome diverser Spezies und den Fortschritten in der Transkriptom-Analyse hat die Erforschung der Evolution des Immunsystems in den letzten Jahren großen Auftrieb bekommen. Die Existenz der adaptiven Immunantwort geht bis auf den Ursprung der Wirbeltiere zurück: bereits Knorpelfische besitzen Immunglobuline, T-Zell-Rezeptoren und alle anderen wesentlichen Bestandteile des adaptiven Immunsystems [Reviews: Eason et al., 2004; Cannon et al., 2004]. Auch die somatischen Diversifikationsmechanismen V(D)J-Rekombination, Klassenwechsel, Genkonversion und somatische Hypermutation sind bereits entwickelt, wenngleich sich der Einfluss dieser Mechanismen und die Struktur der Immunglobulinloci im Verlauf der Evolution verändern. Bei kieferlosen Wirbeltieren gibt es ebenfalls eine adaptive Immunantwort, es konnten jedoch keine Immunglobulin-artigen Antigenrezeptoren gefunden werden. Stattdessen wurden bei den Agnatha variable Lymphozytenrezeptoren mit Leucin-reichen repetitiven Sequenzmotiven gefunden, die offenbar zur Antigenbindung genutzt werden [Pancer et al., 2004; Pancer et al., 2005]. Auch bei Evertebraten gibt es Hinweise auf die Existenz von Mechanismen zur Diversifizierung von Immunrezeptoren [Watson et al., 2005; Reviews: Little et al., 2005; Litman et al., 2005; Flajnik und Du Pasquier, 2004]. Das Prinzip der Rezeptordiversifikation zur Antigenerkennung ist demnach offenbar weit verbreitet; Unterschiede bestehen in der Art der Rezeptormoleküle und der Komplexität der Diversifikation.

Unter den Säugetieren nehmen die Immunglobulin-Gene von Maus und Mensch eine Sonderstellung ein, da die Genkonversion als Diversifikationsmechanismus hier keine Rolle spielt [Review: Flajnik, 2002]. Insofern ist die Maus als Modell zur Untersuchung menschlicher Immundefizienzen, Autoimmunerkrankungen und Infektionsabwehr nicht nur aufgrund der Möglichkeit der Erzeugung von Knock-out-Mäusen, sondern auch wegen der genetischen Ähnlichkeit in besonderem Maße geeignet. Die humanen Immunglobulinloci wurden in den neunziger Jahren intensiv untersucht. Auch Kappa- und Lambda-Locus der Maus sind vollständig sequenziert [Reviews: Matsuda, 2004; Zachau, 2004; Lefranc, 2004]. Der murine Schwerekettenlocus ist im Rahmen des Maus-Genom-Projektes größtenteils, aber nicht vollständig sequenziert worden [Review: Riblet, 2004; Waterston et al., 2002]. Die genomischen Sequenzen von Maus und Mensch stehen über den Ensembl Browser im Internet zur Verfügung [Hubbard et al., 2005]. Im Maus-Genom-Projekt wurde der Stamm C57BL/6 sequenziert, der dem Igh^b-Haplotyp zugeordnet wird [Green, 1979; Brodeur und Riblet, 1984]. Viele Untersuchungen der murinen Immunantwort fanden und finden jedoch an Mäusen oder Zelllinien vom Igh^a-Haplotyp statt, inklusive der intensiven Analysen der konstanten Regionen des Schwerekettenlocus zu Beginn der achtziger Jahre, die an BALB/c-Mäusen vorgenommen wurden [Shimizu et al., 1982; Review: Honjo, 1983]. Da die Immunglobulinloci ausgeprägte Polymorphismen aufweisen, ist die Kenntnis der Haplotypen-spezifischen Sequenz für alle sequenzbasierten Experimente von entscheidender Bedeutung. Die Wahl des Mausstammes ist bei der Durchführung von Infektionsexperimenten keineswegs beliebig, weil beispielsweise die Suszeptibilität für einen Erreger bei verschiedenen Stämmen durchaus sehr verschieden ausfallen kann [Fowell und Locksley, 1999; Else und Wakelin, 1989].

In dieser Arbeit wird die Sequenz des IgH-Locus der 129/Sv-Maus bearbeitet. Die Sequenz-Annotation umfasst zunächst die Positionen der V-, D- und J-Segmente und deren RSS-Elemente sowie die Exone der konstanten Region. Die vorliegende Sequenz umfasst 1382053 bp und deckt neben der C-, D- und J-Region den JH-proximalen Teil der V-Region ab. Die Gesamtgröße des murinen IgH-Locus wird auf 3 Mb geschätzt [Review: Riblet, 2004]. Der 129/Sv-Stamm wird dem Igh^a-Haplotyp zugeordnet und sollte deshalb große Ähnlichkeiten mit BALB/c aufweisen. Embryonale Stammzelllinien vom 129/Sv-Stamm tragen in chimären Tieren besonders effizient zur Keimbahnbesiedlung bei, weshalb der Stamm häufig für transgene Mausexperimente verwendet wird. Die Verfügbarkeit der genomischen Sequenz ist auch Ausgangspunkt für die Untersuchung aller Vorgänge und Mechanismen, die sich im Verlauf der B-Zell-Entwicklung am Schwerekettenlocus abspielen. Die Aufklärung des Schwerekettenlocus von 129/Sv bietet daher eine wichtige Grundlage für verschiedenste experimentelle Ansätze. Weiterhin erlaubt die Verfügbarkeit der Sequenz des Igh^a- und Igh^b-Haplotyps eine detaillierte Analyse der Polymorphismen und verspricht Hinweise auf die Evolution dieses schnell evolvierenden Locus.

1.2 Immunglobulin-Datenbanken und V-Gen-Klassifikation

Seit die ersten Proteinsequenzen von Immunglobulinen in den siebziger Jahren veröffentlicht wurden, wurden sie gesammelt und systematisch klassifiziert. So veröffentlichte die Gruppe von Elvin Kabat 1976 die erste Sammlung in Buchform, “Variable Regions of Immunoglobulin Chains”. Kabat setzte historische Maßstäbe, indem er für die Immunglobuline eine einheitliche Numerierung der Aminosäurepositionen einführte und die Variabilität der einzelnen Positionen bestimmte [Kabat und Wu, 1970]. Durch diese Untersuchungen wurde schnell deutlich, dass es innerhalb der variablen Region drei Bereiche besonders ausgeprägter Variabilität gibt, sogenannte “hyper variable regions” (HVR). Die Aminosäuren dieser Bereiche sind an der Antigenbindung beteiligt, so dass sie auch als “complementary determing regions” (CDR) bezeichnet werden. Die CDRs sind von relativ gut konservierten Bereichen umgeben, den “frame work regions” (FR). In der Struktur der Immunglobulin-Domäne, die die Faltung eines antiparallelen ß-Faltblattes aufweist, befinden sich die FRs innerhalb des Faltblattes, während die CDRs die exponierten Loop-Positionen einnehmen. Kabat sortierte die Immunglobuline zunächst nach FR1-Sequenzen [Kabat et al., 1976]. Renate Dildrop nahm später für die variablen Regionen des murinen Schwerekettenlocus eine Klassifikation vor, die auf dem Vergleich der gesamten Aminosäuresequenz basiert [Dildrop, 1986]. Auf diese Art wurden die murinen Igh-V-Aminosäuresequenzen in sieben Gruppen eingeteilt, die mit den vier von Kabat definierten Gruppen teilweise korrespondieren. Brodeur und Riblet nahmen auf Grundlage der DNA-Sequenzen eine Einteilung in sieben V-Gen-Familien vor, die sehr gut mit der Dildrop-Klassifikation übereinstimmt [Brodeur und Riblet, 1984]. Später wurden weitere V-Gen-Familien entdeckt [Review: Kofler et al., 1992], so dass man heute fünfzehn Familien funktioneller V-Gene im murinen

IgH-Locus kennt. Cyrus Chothia und Arthur Lesk ergänzten die Klassifikation der Immunglobuline auf struktureller Ebene und stellten fest, dass für die CDR-Regionen eine Reihe von Standard-Konformationen existieren, die sie als kanonische Strukturen bezeichneten [Chothia und Lesk, 1987; Al-Lazikani et al., 1997; Morea et al., 1998].

Die Kabat-Datenbank wurde ständig erweitert und blieb stets unabhängig von den großen öffentlichen Nukleotid-Sequenz-Datenbanken GenBank/EMBL/DDBJ. Aufgrund einer Kommerzialisierung ist sie nicht mehr öffentlich zugänglich; die letzte frei verfügbare Version der Sequenzen ist von 1992 (ftp://ftp.ebi.ac.uk/pub/databases/kabat/Rel5.0/). Die zweite große Immunglobulindatenbank, IMGT/LIGM-DB, wurde im Rahmen des “International Immunogenetics Information System” (IMGT) eingerichtet, das 1989 von Marie-Paul Lefranc gegründet wurde [Lefranc et al., 1999] (http://imgt.cines.fr/). Die IMGT/LIGM-Datenbank enthält alle Einträge der EMBL-Bank, die dort als Immunglobuline oder T-Zell-Rezeptoren annotiert sind und wird regelmäßig aktualisiert [Lefranc et al., 2005]. Die qualitative Bewertung der IGMT/LIGM-Einträge wird manuell vorgenommen. So wurden einige Sequenzen mit umfangreicher Annotation versehen und es wurden Listen von Keimbahn-V-Gen-Sequenzen angelegt, die “IMGT reference directory sets”. Die ständige visuelle Bewertung und Aktualisierung der derzeit insgesamt 96000 IMGT/LIGM-Einträge ist jedoch nicht möglich, und es gibt keine Informationen darüber, ob und in welchem Umfang die manuelle Annotation fortgesetzt wird.

Im Rahmen der Sequenzierung der humanen Immunglobulinloci in den neunziger Jahren erstellte Ian Tomlinson eine Liste von humanen Keimbahn-V-Gen-Sequenzen, die Vbase-Datenbank [Cook und Tomlinson, 1995] (http://vbase.mrc-cpe.cam.ac.uk/). In Vbase werden zu jedem Keimbahn-V-Gen

alle V(D)J-Rearrangements aufgeführt, die zur Zeit der Erstellung von Vbase in der GenBank/EMBL-Bank/DDBJ enthalten waren. Vbase ist die umfassendste humane V-Gen-Datenbank, und die Angabe der V(D)J-Rearrangements ist eine wichtige Information für den Anwender. Leider wurde Vbase seit 1997 nicht mehr aktualisiert.

Das NCBI bietet eine Liste von Keimbahn-V-Gen-Sequenzen an, die aus den wichtigsten Veröffentlichungen zu den Immunglobulinloci von Maus und Mensch zusammengestellt wurde (http://www.ncbi.nlm.nih.gov/igblast/showGermline.cgi). Neben diesen “Ig germline genes” gibt es eine “Ig sequence database”, die alle Nucleotidsequenzen mit hinreichender Ähnlichkeit zu den “Ig germline genes” enthält.

Zur Einordnung einer beliebigen experimentell ermittelten Sequenz eines V(D)J-Rearrangements ist eine Anwendung nötig, die diese Sequenz mit den bekannten Keimbahn-V-Gen-Sequenzen vergleicht und eine Analyse der CDR3-Region vornimmt, um die Position und Art des J-Segments und gegebenfalls des D-Segments zu ermitteln. Für den Zugriff auf die IMGT Referenzsequenzen und die Vbase-Datenbank steht das DNAPLOT-Programm zur Verfügung [Müller et al., nicht veröffentlicht] (http://www.dnaplot.org). Basierend auf der Erstellung von Lückenmustern führt DNAPLOT einen Vergleich mit den jeweiligen Keimbahn-Sequenzen durch und gibt als Ergebnis ein Alignment der Anfrage-Sequenz mit den besten Treffer-Sequenzen aus, wobei der V-Region ein Alignment der D- und J-Segmente folgt. Das NCBI bietet für den Sequenzvergleich mit den “Ig germline genes” den “Ig-BLAST” an (http://www.ncbi.nlm.nih.gov/projects/igblast/). Diese modifizierte BLAST-Suche erfüllt einen ähnlichen Zweck wie DNAPLOT, wobei die Bewertung der Übersichtlichkeit und Funktionalität der Ausgaben dem Geschmack des Nutzers

überlassen bleibt. Entscheidend für die Wahl der Anwendung ist primär der zugrunde liegende Keimbahn-V-Gen-Datensatz.

Bei den KeimbahnV-Gensequenzen des IMGT und des NCBI fehlt die Information über Rearrangements und es ist nicht ersichtlich, ob und wann eine Aktualisierung der Verzeichnisse erfolgt. Tatsache ist, dass die beiden Datensätze nicht identisch sind. In den “Ig germline genes” des NCBI wird jeweils nur ein Allel eines V-Gens aufgeführt, und die Auswahl der Publikationen limitiert das präsentierte Repertoire. Die IMGT/LIGM-Datenbank beschränkt sich auf annotierte EMBL-Bank-Sequenzen, so dass unter anderem die gewöhnlich nicht annotierten BAC-Sequenzen und Contigs, die im Rahmen der Genom-Sequenzierprojekte erstellt wurden, nicht erfasst werden. Man kann davon ausgehen, dass diese Beschränkung sich auch auf die “Ig reference sequences” auswirkt. Es existiert demnach keine aktuelle Keimbahn-V-Gen-Sammlung, die sich auf die systematische Untersuchung aller verfügbaren Sequenzen stützt.

Um eine umfassende und aktuelle Liste von murinen und humanen Keimbahn-V-Gen-Sequenzen aus der GenBank/EMBL-Bank/DDBJ zu erstellen, wurde in dieser Arbeit ein Analyse-Prozess entwickelt, der alle Sequenzen der GenBank/EMBL-Bank/DDBJ nach V-Genen durchsucht und diese klassifiziert. Ziel des Prozesses ist eine sequenzbasierte Sortierung und Bewertung der V-Gene, die die Zuordnung von Keimbahn-V-Genen und Rearrangements nach Art der Vbase-Datenbank erlaubt. Der Prozess kann auch zur Analyse großer genomischer Sequenzen verwendet werden, so dass die Einbeziehung der Genom-Sequenzierprojekte möglich ist. Mit Hilfe des Prozesses wird der Datensatz einer dynamischen Keimbahn-V-Gen-Datenbank für Maus und Mensch erzeugt. Diese Datenbank ist, soweit möglich, mit den relevanten bestehenden Datenbanken verbunden worden. Die Automatisierung des

Prozesses ist notwendig, um eine regelmäßige Aktualisierung und die damit verbundene Analyse enormer Datenmengen zu bewerkstelligen. Ein weiterer Vorteil der Automatisierung ist die Möglichkeit der Erweiterung der Datenbank auf andere Spezies und T-Zell-Rezeptor-V-Gene.

Bei der Charakterisierung des Immunglobulinlocus der 129/Sv-Maus hat der V-Gen-Analyse-Prozess einen entscheidenden Anteil an der Annotation der V-Gene. Zum einen kann ein Teil der Annotation auf diese Art automatisiert werden. Zum anderen liefert die Information über Rearrangements wichtige Hinweise auf die Funktionalität der V-Gene. Insgesamt schafft der Prozess eine breite Datenbasis, die eine wichtige Voraussetzung für vergleichende Analysen der Immunglobulinloci von Maus, Mensch und nahe verwandten Spezies liefert.

1.3 Aufgabenstellung

Ziel dieser Arbeit ist die Entwicklung einer dynamischen Keimbahn-V-Gen-Datenbank mit Hilfe eines automatischen V-Gen-Analyse-Prozesses. Dieser Prozess soll die primären Nukleotidsequenzdatenbanken nach V-Gen-Sequenzen durchsuchen, die Sortierung und Evaluierung dieser Sequenzen ermöglichen und Verbindungen zu den Einträgen anderer Datenbanken herstellen. Weiterhin soll der proximale Teil des Immunglobulin-Schwerekettenlocus der 129/Sv-Maus annotiert werden. Der V-Gen-Datensatz der Datenbank soll dabei als Grundlage für die Annotation der variablen Region dienen. Die Arbeit soll eine Basis für vergleichende Untersuchungen der Immunglobulinloci verschiedener Mausstämme und Spezies und Studien zur Evolution der Immunglobuline schaffen.

2. Ergebnisse

2.1 Entwicklung einer dynamischen Keimbahn-V-Gen-Datenbank

Beim Vergleich der V-Gen-Datenbanken Vbase, IMGT/LIGM und Kabat zeichnet sich Vbase durch die Zuordnung von Rearrangements und Keimbahn-V-Genen aus. Dieses Prinzip erlaubt nicht nur die Reduktion der Redundanz der V-Gen-Einträge der EMBL-Bank, sondern liefert auch wichtige Informationen über die Funktionalität eines Keimbahn-V-Gens. Für die Entwicklung der hier beschriebenen Datenbank wurde das Prinzip übernommen und automatisiert. Um dies im Namen der Datenbank deutlich zu machen, wurde sie mit der freundlichen Erlaubnis von Ian Tomlinson, dem Gründer von Vbase, „VBASE2“ genannt. Man kann VBASE2 als eine Weiterentwicklung von Vbase betrachten, weil der Datensatz durch die automatische Generierung dynamisch geworden ist und nicht auf humane V-Gene beschränkt ist.

2.1.1 Automatischer Prozess zur Erkennung und Analyse von Keimbahn-V-Gen-Sequenzen in der EMBL-Bank und in Ensembl

2.1.1.1 Prozessbeschreibung

Es wurde ein Prozess entwickelt, der die Nukleotidsequenzen der EMBL-Bank und der Ensembl-Chromosomen nach V-Genen durchsucht und die neue V-Gen-Datenbank VBASE2 erzeugt. VBASE2 enthält alle Keimbahn-V-Gene, die in den Quelldatenbanken identifiziert werden konnten, und verweist auf die jeweiligen Originaleinträge der Quelldatenbanken. In diesem Sinne dient der Prozess der Sortierung und Evaluierung von V-Gen-Einträgen und V-Gen-Sequenzen in den

Quelldatenbanken. Der Nutzen des Prozesses liegt zum einen in der Reduktion der Redundanz in der EMBL-Bank, da es für die meisten V-Gene mehrere, teilweise über hundert Einträge gibt. Zum anderen ist eine wesentliche Aufgabe des Prozesses die Unterscheidung von Keimbahn-V-Genen gegenüber den somatisch mutierten Sequenzen, die ebenfalls in großer Zahl in der EMBL-Bank vorliegen. Weiterhin dient der Prozess der Annotation der V-Gene in Ensembl.

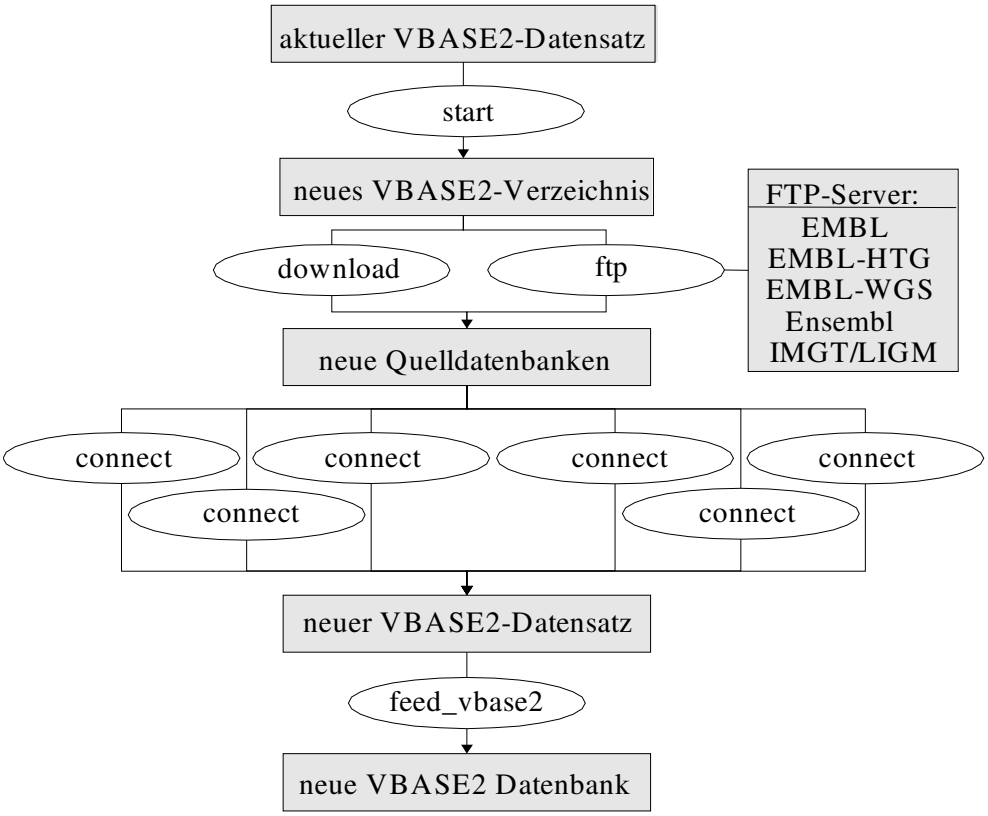
Die Automatisierung des Prozesses ist notwendig, da die regelmäßige Aktualisierung der Quelldatenbanken die entsprechende Anpassung des V-Gen-Datensatzes erfordert. Der Prozess wurde für die jeweils drei Immunglobulinloci von Maus und Mensch entwickelt. Er ist jedoch auch auf andere V-Gen-Loci anwendbar, sofern diese eine ähnliche Struktur aufweisen und eine ausreichende Menge an V-Gen-Sequenzen bekannt ist.

Der Ablauf des Prozesses ist in Abbildung 2.1 dargestellt. Zunächst werden alle benötigten Daten aus dem Internet lokal gespeichert. Aus den Quelldatenbanken EMBL-Bank und Ensembl werden nur die speziesspezifischen Subdatensätze genutzt. Weiterhin werden aus der EMBL-Bank die Datensätze “HTG” (High Throughput Genomic Sequences) und “WGS” (Whole Genome Shotgun Sequences) heruntergeladen. Diese großen Datensätze sind nicht Bestandteil des Standard-Spezies-Datensatzes, enthalten jedoch zahlreiche V-Gen-Sequenzen. Neben den externen Sequenzdatenbanken werden interne Daten benötigt, wie beispielsweise Informationen über V-Gen-Familien und die V-Gen-Sequenzen aus dem aktuellen VBASE2-Datensatz. Die anschließende Analyse der Quelldatenbanken erfolgt für jeden zu untersuchenden V-Gen-Locus einzeln. Da der Prozess auf einem Linux-Cluster mit 16 Knoten durchgeführt wird, können die sechs Analyse-Prozesse zur Untersuchung der

Immunglobulin-V-Gene von Maus und Mensch parallelisiert werden. Am Ende des Prozesses wird automatisch eine neue PostgresQL Datenbank erzeugt, in der der neue Datensatz gespeichert wird.

Abbildung 2.1: Übersicht über den VBASE2-Generationsprozess

Dargestellt sind Daten (grau hinterlegte Kästen) und die Namen der darauf zugreifenden Perl-Skripte (Elipsen). Die Aufgaben der Skripte werden in Tabelle 2.1 erklärt.



Grundlage des Analyse-Prozesses sind das NCBI-blastall-Programm und DNAPLOT. Die BLAST-Suche von V-Genen des aktuellen Datensatzes in den Quelldatenbanken dient der Erkennung von potentiellen Keimbahn-V-Genen. Diese V-Gen-Kandidaten werden mit Hilfe des DNAPLOT-Programms sortiert und analysiert. DNAPLOT ist ein Programm zum Alignment von V-Gen-Sequenzen, das zusätzlich zahlreiche Möglichkeiten zur DNA-Sequenzanalyse

und -formatierung bietet. Der Prozessablauf ist in Perl programmiert.[1]

Tabelle 2.1: Aufgaben der Perl-Skripte im VBASE2-Generationsprozess

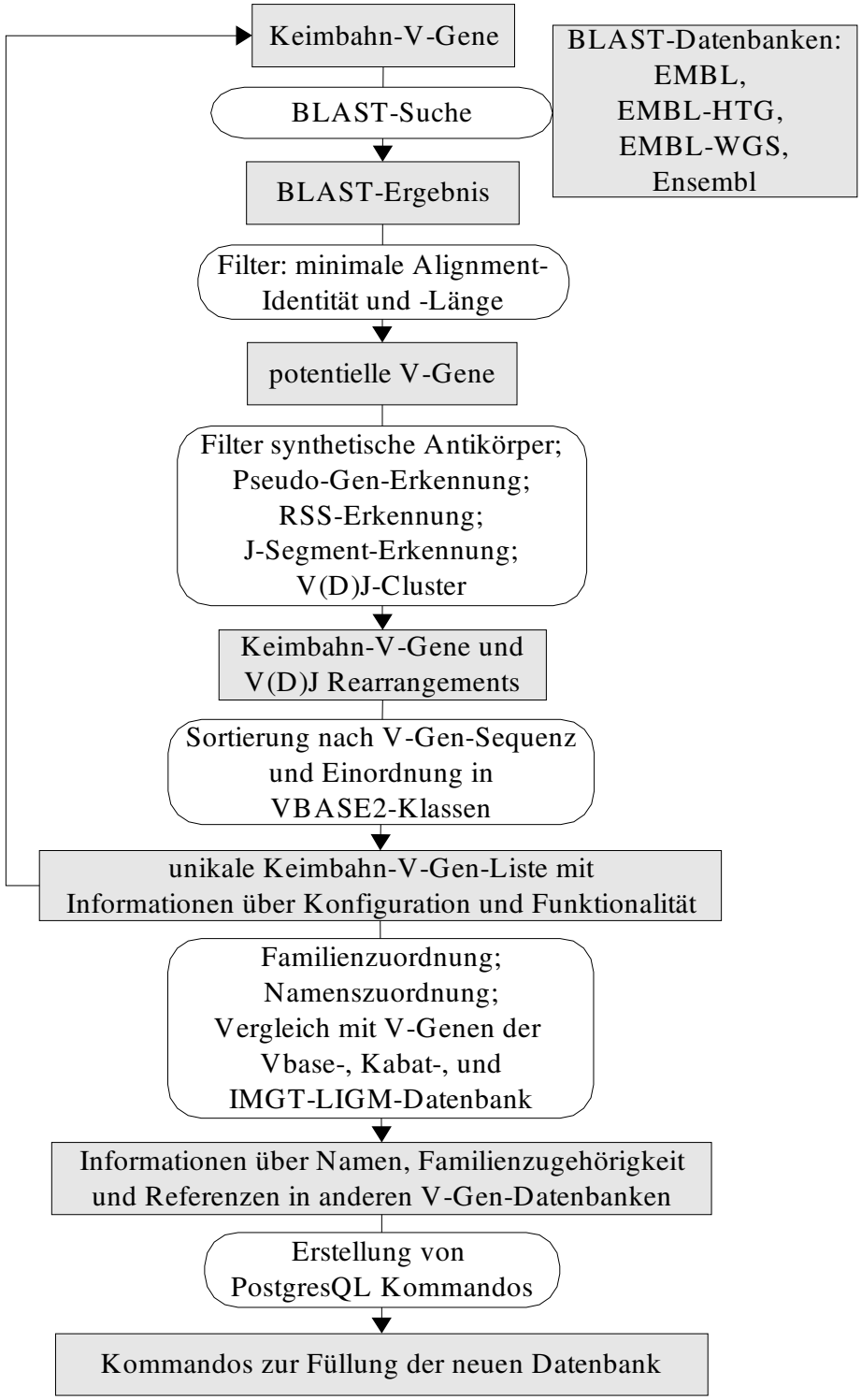
Perl-Skript	Aufgabe
start	Kopiert alle Daten des aktuellen Datensatzes, die für den neuen Prozess benötigt werden, in ein neues Arbeitsverzeichnis.
download	Bereitet den ftp-Zugriff auf die Daten aus dem Internet vor; formatiert die heruntergeladenen Sequenzen und erstellt Datenbanken für die BLAST-Suche.
ftp	Speichert Daten aus dem Internet über ftp auf dem lokalen Rechner.
connect	Ruft alle Skripte auf, die für den locusspezifischen Teil des Prozesses nötig sind: - V-Gen-Erkennung und -Analyse - Suche von V-Genen in den Datensätzen von anderen V-Gen-Datenbanken - Vergleich des neuen Datensatzes mit dem aktuellen Datensatz - Erstellung von PostgresQL-Kommandos zur Füllung der Datenbank
feed_vbase2	Erzeugt eine PostgresQL-Datenbank und fügt die im Prozess erzeugten Daten ein.

Abbildung 2.2 zeigt die automatische V-Gen-Analyse, mit der die in den Quelldatenbanken vorhandenen V-Gene eines Immunglobulinlocus sortiert und klassifiziert werden können. Dabei ist an allen V-Gen-spezifischen Untersuchungen das DNAPLOT-Programm wesentlich beteiligt. Die Perl-Skripte, die den Prozessablauf steuern, sind in Tabelle 2.1 aufgeführt.

1 Der Wortlaut der Perl-Skripte kann in dieser Arbeit nicht abgedruckt werden, da eine mögliche kommerzielle Verwertung des VBASE2-Generationsverfahrens derzeit durch die Firma Ascenion GmbH geprüft wird.

Abbildung 2.2: Ablauf der automatischen V-Gen-Analyse

Die Abbildung zeigt den Teil des Prozesses, der durch das connect-Skript kontrolliert wird. Daten sind in grau hinterlegten Kästen dargestellt, die Schritte der Analyse in abergundeten weißen Kästen. An allen Schritten mit Ausnahme der BLAST-Suche ist das DNAPLOT-Programm beteiligt.



Die in der BLAST-Suche ermittelten V-Gen-Kandidaten aus der EMBL-Bank werden zunächst auf synthetische Antikörper geprüft. Da bei synthetischen Antikörpern nicht sichergestellt ist, dass die V-Gen-Sequenz auch *in vivo* vorkommt, werden sie von der Analyse ausgeschlossen. Der Filter basiert erstens auf der Suche nach typischen Textbausteinen wie „ScFv“ und „dsFv“ in den EMBL-Bank-Einträgen. Zweitens wird geprüft, ob sich hinter der identifizierten V-Gen-Sequenz noch ein weiteres V-Gen anschließt. Die Kodierung von schwerer und leichter Kette als Fusionsprotein in einer gemeinsamen kodierenden Sequenz ist typisch für synthetische Antikörper [Review: Rohrbach et al., 2003] und kommt unter natürlichen Bedingungen nicht vor. Drittens werden die EMBL-IDs bekannter synthetischer Antikörper in einer manuell erstellten Liste gespeichert und von der Analyse ausgeschlossen.

Zur Identifizierung von Pseudogenen werden alle V-Gen-Treffer der BLAST-Suche nach Stop-Codons innerhalb der kodierenden Sequenz untersucht. V-Gene, die Stop-Codons enthalten, werden als Pseudogene klassifiziert. Bei V-Genen aus dem Ensembl-Datensatz wird die Nummer des Chromosoms überprüft. Stimmt sie nicht mit dem Chromosom überein, auf welchem sich der jeweilige Immunglobulinlocus befindet, wird das V-Gen als Orphan markiert.

Anschließend werden die V-Gene nach der Konfiguration sortiert: V(D)J Rearrangements werden durch Suche nach J-Segmenten am 3'-Ende von V-Genen aus der EMBL-Bank identifiziert. Im gleichen Sequenzbereich der EMBL-Bank-V-Gene wird nach RSS-Elementen gesucht. Bei Erkennung eines RSS-Elements erfolgt die Zuordnung als Keimbahn-konfiguriertes V-Gen. Für alle V-Gene aus Ensembl, EMBL-HTG, EMBL-WGS oder BAC-Sequenzen aus der EMBL-Bank wird die Keimbahn-Konfiguration aufgrund der genomischen Herkunft angenommen. Sequenzen, bei denen die Konfiguration nicht bestimmt werden

kann, werden von weiteren Untersuchungen ausgenommen.

In einem besonderen Fall werden auch rearrangierte V-Gene als Belege für Keimbahn-Sequenzen genutzt: wenn ein V-Gen in zwei unabhängigen Rearrangements gefunden wird, ist dies ein deutlicher Hinweis darauf, dass in diesem V-Gen keine somatische Mutation vorliegt. Für das in Abbildung 2.2 als 'V(D)J-Cluster' bezeichnete Verfahren werden alle Rearrangements in Gruppen (Cluster) identischer V-Gen-Sequenzen sortiert, so dass eine Liste von nicht redundanten V-Gen-Sequenzen entsteht. Derartige 'nur einmal vorkommenden' Sequenzen werden im Folgenden als unikal bezeichnet. Besteht eine Gruppe aus mehr als einer Sequenz, werden die (D)J-Bereiche der Rearrangements miteinander verglichen. Gibt es zu einem unikalen V-Gen zwei oder mehr verschiedene Rearrangements, erfolgt die Einstufung als Keimbahn-V-Gen-Sequenz.

Im nächsten Schritt des Prozesses werden die Keimbahn-konfigurierten V-Gensequenzen mit den V(D)J-Rearrangements und mit sich selbst verglichen. Auf diese Art werden unikale Keimbahn-V-Gene mit dazugehörigen Rearrangements ermittelt. Für jede unikale Sequenz wird eine Liste aller Quelldatenbank-Einträge erstellt, die diese Sequenz belegen. Für die Zuordnung eines Rearrangements zu einer Keimbahn-Sequenz ist dabei eine bis zum Ende von FR3 hundertprozentig identische V-Gen-Sequenz erforderlich. Diese Stringenz schließt zwar die Berücksichtigung von somatischen Mutationen bei der Beurteilung der potentiellen Funktionaliät vollständig aus. Sie ist jedoch nötig, um in einem automatischen Prozess Falsch-Positive zu vermeiden.

Um die Qualität einer Keimbahn-V-Gen-Sequenz zu dokumentieren, werden die V-Gene anschließend in drei VBASE2-Klassen eingeordnet:

Klasse 1 enthält V-Gene, die sowohl in rearrangierter Form, als auch in der Keimbahn-Konfiguration gefunden werden. Durch den mindestens doppelten Beleg der Sequenz sind V-Gen-Sequenzen dieser Klasse als besonders zuverlässig einzustufen. Außerdem ist ein V(D)J- Rearrangement ein starker Hinweis darauf, dass das beteiligte V-Gen tatsächlich für einen Antikörper kodiert.

Klasse 2 enthält V-Gene, die nur in der Keimbahn-Konfiguration gefunden wurden. Diese Klasse enthält vor allem V-Gene, die sehr selten oder gar nicht rearrangieren. Viele V-Gene der Klasse 2 sind Pseudogene. Auch fehlerhaft sequenzierte V-Gene werden in diese Klasse einsortiert. Dies betrifft möglicherweise einige Sequenzen, die nur durch eine HTG- oder WGS-Sequenz belegt sind, welche aus nicht abgeschlossenen Sequenzierprojekten stammen, oder auch ältere genomische V-Gen-Sequenzen. Weiterhin besteht die Möglichkeit, dass ein Keimbahn-V-Gen als Rearrangement nicht gefunden wird, weil alle bisher sequenzierten Rearrangements somatisch mutiert vorliegen. Solche Sequenzen können ebenfalls in Klasse 2 enthalten sein.

Klasse 3 enthält V-Gene, die nicht in der Keimbahn-Konfiguration gefunden wurden, sondern durch die Bildung von V(D)J-Clustern als Keimbahn-V-Gene identifiziert wurden. Dies sind möglicherweise Allele, deren genomische Sequenz bisher noch nicht untersucht wurde.

Nach dieser Klassifikation werden die unikalen V-Gene weiter untersucht. Zum einen wird ein Vergleich mit Familien-Konsensus-Sequenzen durchgeführt, der

die Zuordnung zu einer V-Gen-Familie ermöglicht. Zum anderen werden V-Gen-Namen aus der Literatur den entsprechenden Sequenzen zugeordnet. Da teilweise mehrere V-Gen-Nomenklaturen parallel verwendet werden, werden hier alle Namen, die zu einem V-Gen bekannt sind, gespeichert. Schließlich werden die V-Gene mit DNAPLOT in der V-Gen-Datenbank von Kabat, in der Vbase-Datenbank sowie in der IMGT/LIGM-Datenbank gesucht und die Treffer gespeichert.

Die Sequenzen der V-Gene werden mit allen Informationen, die im Verlauf des Prozesses ermittelt wurden, in Kommandos eingefügt, die die Füllung der anschließend erzeugten Datenbank ermöglichen. Am Ende des Prozesses steht also eine Reihe von Textdateien, die in die neu erzeugte Datenbank eingelesen werden können.

In einem erneuten Prozessablauf werden die VBASE2-V-Gen-Sequenzen für die BLAST-Suche als Vergleichssequenzen benutzt. Auf diese Art kann das Prozessergebnis mit jeder Aktualisierung verbessert werden, da das V-Gen-Spektrum sukzessive erweitert werden kann. Dies ist insbesondere nützlich für die Anwendung des Prozesses auf Spezies, bei denen bisher wenig Keimbahn-V-Gene annotiert sind (siehe auch: 2.1.3.1).

2.1.1.2 Prozessergebnis

Der im Juni 2005 aktuelle Datensatz von VBASE2 (Tabelle 2.2) umfaßt 1129 unikale Keimbahn-V-Gene und -Pseudogene der Immunglobulinloci von Maus und Mensch, die durch insgesamt 7008 Referenz-Sequenzen in der EMBL-Bank und in Ensembl belegt sind.

Für 347 von 1089 keimbahn-konfigurierten Sequenzen wurde mindestens ein Rearrangement ermittelt (Klasse 1). Die restlichen 742 Keimbahn-Konfigurationen gehören zu Klasse 2, 380 davon tragen ein Stop-Codon in der Sequenz. 210 Klasse-2-Sequenzen konnte keine V-Gen-Familie zugeordnet werden. Dies ist ein Hinweis darauf, dass es sich um verkürzte V-Gene oder V-Gen-Relikte handelt, deren Sequenz sich um mehr als 70% von den funktionellen V-Genen unterscheidet. 51 Klasse-2-Sequenzen werden auf anderen Chromosomen lokalisiert, so dass sie als V-Gen-Orphans bezeichnet werden. 43 Orphans werden auf humanen Chromosomen detektiert, während im murinen Genom nur elf Orphans gefunden werden. 40 V-Gene wurden aufgrund der Identifizierung verschiedener Rearrangements des gleichen V-Gens als Keimbahn-V-Gene ausgewiesen, obwohl die Sequenz in der Keimbahn-Konfiguration bisher nicht bekannt ist (Klasse 3).

Die Verteilung der Sequenzen auf die drei VBASE2-Klassen ist bei den insgesamt sechs Loci sehr ähnlich (Tabelle 2.2). Etwa 30% der VBASE2-V-Gen-Einträge gehören zur Klasse 1. Der Anteil der Klasse-3-Sequenzen beträgt weniger als 5%.

Tabelle 2.2: Anzahl der V-Gen-Einträge in VBASE2

Datensatz vom Juni 2005

Spezies, Locus	Klasse 1	Klasse 2	Klasse 3	Summe
Mensch, IGHV	57	205	8	270
Mensch, IGKV	44	97	5	146
Mensch, IGLV	45	77	6	128
Maus, IGHV	123	239	17	379
Maus, IGKV	75	122	4	201
Maus, IGLV	3	2	0	5
Summe	347	742	40	1129

2.1.1.3 Prozessvalidierung

2.1.1.3.1 Parameter-Optimierung

Die automatische V-Gen-Analyse erfordert eine große Anzahl von Parametern oder Grenzwerten, die in den einzelnen Programmen eingestellt werden müssen. Die wichtigsten Parameter sollen in diesem Abschnitt genannt und einige Beispiele detailliert dargestellt werden.

Bei der Optimierung der Parameter stehen sich meist Qualität und Quantiät gegenüber: Einerseits muss die Richtigkeit des Prozessergebnisses gewährleistet sein, das bedeutet: Es muss sichergestellt werden, dass der VBASE2-Datensatz ausschließlich Keimbahn-V-Gene enthält. Andererseits sollen möglichst alle V-Gene der Quelldatenbanken von der Analyse erfasst werden und in den VBASE2-Datensatz eingehen. In einigen Fällen spielen auch Rechenzeit und Datenmengen eine Rolle, wie zum Beispiel bei der Einstellung der BLAST-Parameter. Die Parameter wurden stets so gesetzt, dass keine falsch-positiven Ergebnisse zu erwarten sind. So ist beispielsweise für die Zuordnung eines Rearrangements zu einer Keimbahn-Sequenz eine hundertprozentige Übereinstimmung der V-Gen-Sequenz erforderlich (siehe auch 2.1.1.1.), so dass somatische Mutationen von der Betrachtung vollständig ausgeschlossen werden. Zahlreiche V-Gen-Einträge der EMBL-Bank können im Prozess nicht als Referenzen berücksichtigt werden, weil die Sequenz direkt hinter dem V-Gen-Ende abbricht und daher die Konfiguation nicht bestimmt werden kann. Insgesamt erfordert die Automatisierung eine höhere Stringenz als für eine manuelle Untersuchung der V-Gene nötig wäre. Im Gegenzug ermöglicht die Automatisierung die Bearbeitung von Datenmengen, die manuell keinesfalls geleistet werden könnte.

Grundlage des Prozesses ist die Charakterisierung unbekannter Sequenzen aufgrund von Ähnlichkeit mit bekannten Sequenzen. Dabei spielt die Art der bekannten Sequenzen eine ebenso wichtige Rolle wie die minimale Identität, die für die Anerkennung der Ähnlichkeit erforderlich ist. Entscheidenden Einfluss haben weiterhin die Position und die Länge des Alignments sowie programmspezifische Parameter wie beispielsweise die Anzahl der auszugebenden Sequenzen beim blastall-Programm.

Die Tabelle 2.3 zeigt eine Übersicht über die wichtigsten numerisch erfassbaren Parameter mit den ermittelten optimalen Werten. Eine besondere Bedeutung kommt außerdem den sogenannten Master-Sequenzen zu, die für das V-Gen-Alignment und für die Familienzuordnung benötigt werden. Die Master-Sequenzen erlauben die einheitliche Nummerierung der Nukleotide innerhalb der V-Gene (siehe Methoden: DNAPLOT). Die V-Gen-Master wurden von Werner Müller zur Verfügung gestellt und im Rahmen dieser Arbeit nicht bearbeitet. Die Sequenzen der J-Segmente wurden von der IMGT-Webseite übernommen.

Tabelle 2.3: Parameter im V-Gen-Analyse-Prozess des murinen IgH-Locus

Positionsangaben innerhalb der V-Gen-Sequenz oder des Rearrangements beziehen sich auf die IMGT-Nummerierung [Lefranc et al., 2003].

Programm	Aufgabe	Parameter	optimaler Wert
blastall	BLAST-Suche der V-Gene in den Quelldatenbanken	Anzahl der auszugebenden Alignments (b)	5000
sortblast	Sortierung der BLAST-Ergebnisse	minimale Alignment-Identität	80 %
		minimale Alignment-Länge	200 bp
filter_synth	Filter für synthetische Antikörper	minimale Identität mit bekanntem V-Gen	65 %
		minimale Abdeckung des zweiten V-Gen-Bereichs	50 bp
find_J	Erkennung von J-Segmenten	minimale Abdeckung des J-Segment-Bereichs	10 bp
		Bereich, in dem das J-Segment gesucht wird (IMGT-Nummerierung)	Position 330-375
		minimale Identität mit bekanntem J-Segment	90 %
cluster	Bildung von Clustern gleicher V-Gen-Sequenz mit verschiedenen Rearrangements	DNAPLOT-Parameter -comp: Anzahl auszugebender Vergleichsergebnisse	500
cluster	Bildung von Clustern gleicher V-Gen-Sequenz mit verschiedenen Rearrangements,	Ende des zu vergleichenden V-Gen-Bereichs im Rearrangement	Position 313
Vlength	Festlegung des Endes der betrachteten V-Gen-Sequenzen	Ende des zu vergleichenden V-Gen-Bereichs	IGHV: Position 319 IGKV: Position 331 IGLV: Position 334
compare	interner Vergleich aller V-Gen-Sequenzen	Beginn der minimalen Abdeckung	Position 79
		minimale Identiät der V-Gen-Sequenzen für die Zuordnung	100%
get_fam	Zuordnung von V-Gen-Familien	minimale Identität mit Familien-Master	70%

Alle im Folgenden erläuterten Parameter-Beispiele beziehen sich auf V-Gene des Immunglobulin-Schwerekettenlocus der Maus.

Die BLAST-Input-Sequenzen dienen als Vorlagen zur Erkennung eines V-Gens in der EMBL-Bank und in Ensembl und sind deshalb von erheblicher Bedeutung für die Prozessvalidierung. Dabei ist einerseits die Frage zu klären, wie viele verschiedene V-Gen-Sequenzen nötig sind, um die maximale Anzahl an V-Genen zu detektieren. Andererseits spielt die Art der Input-Sequenzen eine Rolle: Sollen nur Klasse1-V-Gene verwendet werden, oder sind auch Pseudo-Gene als Input sinnvoll?

Zur Bewertung des Einflusses der Input-Sequenzen wurde der Prozess mit verschiedenen Sets von Input-Sequenzen durchlaufen und das Ergebnis verglichen (Tabelle 2.4, Abbildung 2.3). Zunächst ist es naheliegend, alle V-Gene der VBASE2-Klasse 1 als Input zu verwenden, weil diese als gesicherte Keimbahn-Sequenzen und mit großer Wahrscheinlichkeit als funktionell einzustufen sind. Um die Grenzen des Prozesses auszuloten, wurde jedoch auch das V-Gen v186.2, prominenter Repräsentant der großen J558-Familie, als einzelne Input-Sequenz getestet. Ebenso wurden die fünfzehn V-Gen-Master-Sequenzen, die alle V-Gen-Familien des murinen Schwerekettenlocus repräsentieren, als Input eingesetzt. Weiterhin wurden die V-Gene der VBASE2-Klasse 2 mit den Klasse-1-V-Genen gemeinsam als Input verwendet. Auch V-Gene der Klasse 3 wurden als Input getestet.

Tabelle 2.4: Einfluss der BLAST-Input-Sequenzen auf die Anzahl der resultierenden VBASE2-V-Gene am Beispiel des Schwerekettenlocus der Maus

Die Tabelle zeigt das Ergebnis des VBASE2-Generationsprozesses auf der Basis von BLAST-Suchen in der EMBL-Bank, Release 88 (Januar 2004).

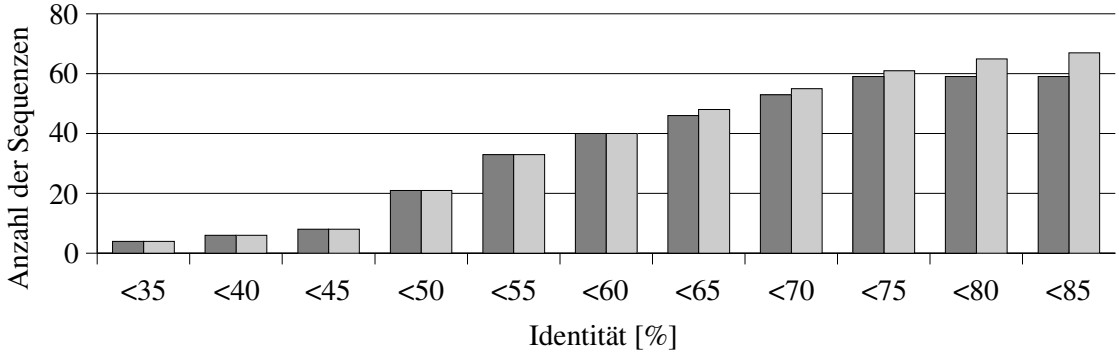
Input		Anzahl der Output-Sequenzen			
Sequenzen	**Anzahl**	**Klasse 1**	**Klasse 2**	**Klasse 3**	**Summe**
v186.2	1	32	58	7	97
V-Gen-Master	15	104	137	12	253
VBASE2 Klasse 1	123	123	181	17	321
VBASE2 Klasse 2	240	119	240	17	376
VBASE2 Klasse 3	17	87	127	16	247
Klasse 2-Sequenzen, die nur durch Klasse 2 gefunden werden	59	0	67	0	67

Die Tabelle 2.4 zeigt, dass mit v186.2 als Suchsequenz nur etwa 30% der V-Gene gefunden werden, die bei der Suche mit den V-Genen der Klasse 1 gefunden werden. Diese sind, sofern es sich nicht um Pseudogene ohne Familienzuordnung handelt, ausschließlich J558-Sequenzen. Dagegen werden fast 80% der mit VBASE2-Klasse 1 gefundenen V-Gene auch mit den V-Gen-Mastern als Input identifiziert. Die Verwendung der Master-Sequenzen als Input erzielt also ein gutes Ergebnis, das durch den Einsatz aller Klasse-1-V-Gene aber noch gesteigert werden kann. Werden die V-Gene der Klasse 2 eingesetzt, die potentielle Pseudo-Gene enthält, so steigt der Anteil an Klasse-2-Sequenzen im Output erheblich. Mit Klasse 2 werden 59 zusätzliche Sequenzen gefunden, die ausschließlich selbst der Klasse 2 angehören. Um zu untersuchen, ob es sich bei diesen Sequenzen tatsächlich um V-Gene oder V-Gen-Relikte handelt, wurde ein Vergleich mit den Klasse-1-Sequenzen durchgeführt und die prozentuale Identität mit dem ähnlichsten Klasse-1-V-Gen bestimmt (Abbildung 2.3). Des weiteren wurde ein erneuter Prozessdurchlauf mit den 59 Klasse-2-

Sequenzen vorgenommen. Das Ergebnis besteht aus 67 Sequenzen, die die 59 Input-Sequenzen enthalten und weiterhin 8 Sequenzen, die bereits durch Klasse-1-Sequenzen gefunden wurden, es enthält jedoch keine neuen Klasse-2-Sequenzen. Die Verwendung von Klasse-2-Sequenzen als Input bewirkt also keine Entfernung der Output-Klasse-2-Sequenzen von den Klasse-1-Sequenzen. Die Verwendung von Klasse 1 und Klasse 2 als Input erzielt demnach eine wertvolle Erweiterung des Outputs und ist somit angebracht. Dagegen bringt der Einsatz von Klasse-3-Sequenzen keine Erweiterung des Outputs und ist deshalb nicht erforderlich.

Abbildung 2.3: Vergleich der Sequenzen, die nur durch Klasse-2-Sequenzen gefunden werden, mit Klasse-1-Sequenzen

59 Klasse-2-Sequenzen, die nur durch Verwendung von Klasse-2-V-Genen als Input detektiert werden, wurden als Input für einen erneuten Prozessdurchlauf verwendet. Der Output umfasst 67 Sequenzen. Input und Output wurden mit V-Genen der Klasse 1 verglichen und die prozentuale Identität zum ähnlichsten Klasse-1-V-Gen bestimmt. Das Diagramm zeigt für Input (dunkelgrau) und Output (hellgrau) die Anzahl der Sequenzen mit einer bestimmten Identität zum ähnlichsten Klasse-1-V-Gen.



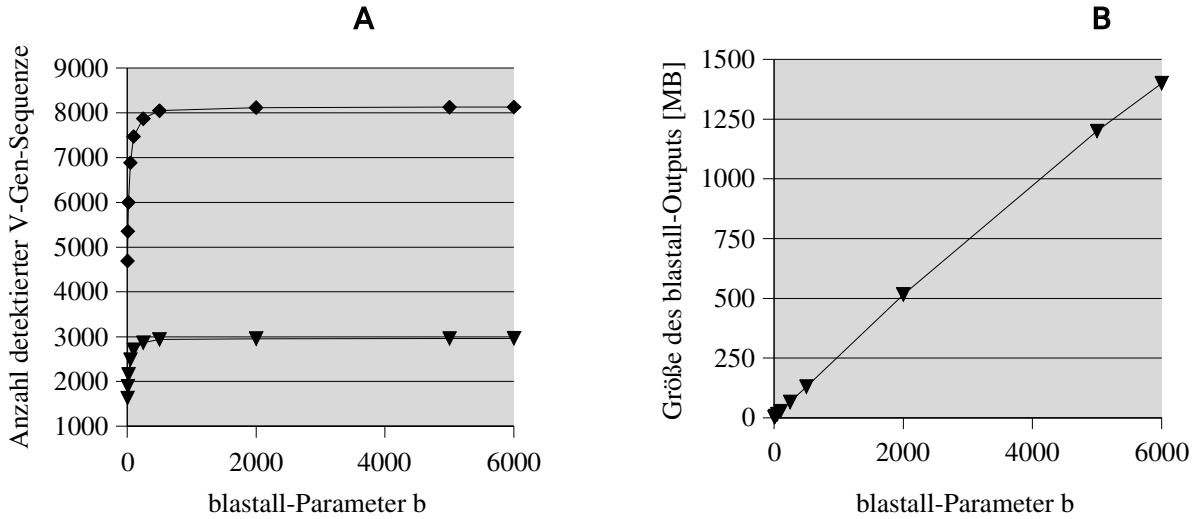
Ebenso wichtig wie die Input-Sequenzen sind bei der BLAST-Suche die Anzahl der auszugebenden Sequenzen, die sich mit dem blastall-Parameter „b“ einstellen lassen. Einige V-Gene sind in der EMBL-Bank hoch redundant vertreten. So ergibt beispielsweise die BLAST-Suche von v186.2 in der EMBL-Bank über 500 Ergebnis-Treffer mit einer Alignment-Länge von mindestens

290 bp und 100% Sequenz-Identität. Um Sequenzen zu finden, die der Such-Sequenz ähnlich sind, muss die Anzahl der Ausgabe-Sequenzen deutlich über der Anzahl der mit der Suchsequenz identischen Einträgen liegen. Andererseits kann die Anzahl der Ausgabe-Sequenzen nicht beliebig erhöht werden, da die Ausgabe nicht relevanter Treffer den verfügbaren Speicherplatz unnötig belastet und die Bearbeitung des BLAST-Ergebnisses verlangsamt. Die Größe der blastall-Output-Datei steigt in dem untersuchten Bereich linear mit steigendem Parameter b (Abbildung 2.4.B). Im Sinne einer Prozessoptimierung muss b deshalb so niedrig wie möglich, aber so hoch wie zur Ausgabe aller relevanten Treffer nötig gewählt werden. Die Abbildung 2.4.A zeigt, dass die Abhängigkeit der detektierten V-Gene von b als Sättigungskurve verläuft. Im Fall des murinen und humanen Schwerekettenlocus wurden 5000 auszugebende Sequenzen als geeigneter Wert für b ermittelt.

Abbildung 2.4: Einfluss der Anzahl der auszugebenden Sequenzen bei der BLAST-Suche (blastall-Parameter b)

A: Dargestellt ist die Anzahl der detektierten V-Gen-Sequenzen bei variablem Parameter b. Rauten: Anzahl der BLAST-Treffer mit mindestens 200 bp Alignment-Länge und 80 % Identität; Dreiecke: Anzahl der durch den Analyse-Prozess detektierten VDJ-Rearrangements.

B: Dargestellt ist die Größe der blastall-Ausgabe-Datei in Abhängigkeit vom Parameter b.



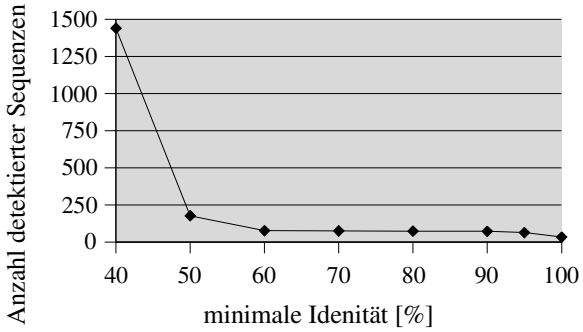
Bei der Bewertung von Sequenzen aufgrund ihrer Ähnlichkeit mit Vorlage-Sequenzen ist der Wert für die minimale Sequenzidentität, die für eine positive Bewertung erfoderlich ist, entscheidend für die Qualität des Prozessergebnisses: Liegt der Grenzwert zu hoch, werden Positive nicht erkannt (Falsch-Negative), bei zu niedriger Schwelle werden Negative als positiv eingestuft (Falsch-Positive). Bei Anwendung des DNAPLOT-Programms spielt außerdem die Position, in welcher der Sequenzvergleich beginnt, eine wichtige Rolle. Im Folgenden soll am Beispiel der V-Gen-Erkennung des Filters für synthetische Antikörper erläutert werden, wie der Wert für die minimale Sequenzidentität festgelegt wurde. Die Bestimmung der optimalen Startposition für den Vergleich wird am Beispiel der J-Segment-Erkennung dargestellt.

Die Erkennung synthetischer Antikörper auf Sequenzebene basiert auf der Tatsache, dass bei den sogenannten „scFv“-Fragmenten schwere und leichte Kette in einer gemeinsamen kodierenden Sequenz zusammengefasst werden [Rohrbach et al., 2003; Filpula und McGuire, 1999]. Wird also in der Sequenz stromabwärts des bereits identifizierten V-Gens ein weiteres V-Gen erkannt, wird diese Sequenz von weiteren Untersuchungen ausgeschlossen. Dabei wird bei den V-Genen des Schwerekettenlocus nach V-Genen des Lambda- und Kappa-Locus gesucht, und umgekehrt bei den Leichtekettenloci nach V-Genen des IgH-Locus. Die Abbildung 2.5 zeigt für den murinen IgH-Locus die Anzahl der erkannten Kappa- und Lambda-V-Gene bei variablen Werten für die minimale Sequenzidentität. Im Bereich zwischen 60% und 90% Identität ist das Ergebnis annähernd konstant. Bei einem Grenzwert von weniger als 50% Identität nimmt die Anzahl der für positiv befundenen Sequenzen stark zu, die Erkennung wird offensichtlich unspezifisch. Bei einem Grenzwert von 65% wurden die für negativ befundenen Sequenzen manuell auf das Vorhandensein

von Falsch-Negativen geprüft, indem die EMBL-Bank-Einträge nach Schlagworten wie „ScFv“ durchsucht wurden. Es wurden jedoch keine Falsch-Negativen gefunden, so dass 65% als minimal erforderliche Sequenz-Identität für die V-Gen-Erkennung übernommen wurde.

Abbildung 2.5: Erkennung von Leichte-Kette-V-Genen in potentiell synthetischen Sequenzen

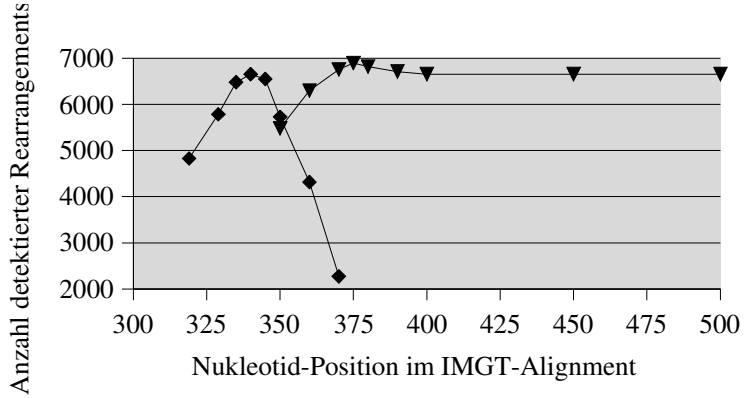
Der automatische V-Gen-Analyse-Prozess identifiziert synthetische Sequenzen durch Vergleich des 3'-Endes von IgH-V-Gen-Sequenzen mit V-Gen-Sequenzen des Kappa- und des Lambda-Locus. Bei dem Sequenzvergleich ist eine minimale prozentuale Identität nötig, um eine Sequenz als synthetisch auszuweisen. Die Abbildung zeigt die Anzahl der detektierten synthetischen Sequenzen in Abhängigkeit vom Parameter 'minimale Identität'.



Zur Identifizierung von V(D)J-Rearrangements wird der Bereich hinter dem V-Gen mit Sequenzen von J-Segmenten verglichen. Dabei hat die genaue Position, in der der Vergleich beginnt und endet, erheblichen Einfluss auf das Ergebnis. Zur Bestimmung der optimalen Position wurden Start- und Endposition des Vergleichs innerhalb des Sequenzbereichs variiert, in welchem das J-Segment bei einem Rearrangement zu erwarten ist [Lefranc et al., 2003]. Das in Abbildung 2.6 dargestellte Ergebnis zeigt für die Start-Position ein Maximum in Position 340 und für die End-Position ein Maximum in Position 375. Diese Positionen wurden als Parameter-Werte in den Prozess übernommen.

Abbildung 2.6: Erkennung von J-Segmenten in V(D)J-Rearrangements

Zur Erkennung von J-Segmenten in V(D)J-Rearrangements mit dem DNAPLOT-Programm ist die Wahl der Region, in welcher das J-Segment gesucht werden soll, nötig. Dargestellt ist die Anzahl der detektierten Rearrangements bei verschiedenen Start- und Endpositionen des J-Segment-Vergleichs. Positionsangaben beziehen sich auf die IMGT-Nummerierung [Lefranc et al., 2003]. Rauten: End-Position 450, Start-Position variabel; Dreiecke: Start-Position 340, Endposition variabel.



2.1.3.2 Validierung des automatisch erzeugten V-Gen-Datensatzes

Um die Vollständigkeit und die Qualität des VBASE2-Datensatzes zu bewerten, wurden die VBASE2-V-Gene mit den V-Genen anderer Keimbahn-V-Gen-Sammlungen verglichen. Neben der Vbase-Datenbank wurden hier auch die 'ABG germline gene directories of the mouse' von Juan Carlos Almagro [Almagro et al., 1997] einbezogen. Almagro bietet auf seiner Webseite eine Sammlung von murinen Keimbahn-V-Gen-Sequenzen an, die aus der IMGT/LIGM-Datenbank, GenBank und aus der Literatur zusammengetragen wurden. Weiterhin wurde der VBASE2-Datensatz mit den 'IMGT reference sequences' [Lefranc et al., 1999] für Maus und Mensch verglichen. Diese Keimbahn-V-Gen-Sammlung des IMGT wurde aus der Literatur und der EMBL-Bank erstellt und soll alle bekannten V-Gen-Allele umfassen. Die Vbase-, ABG- und IMGT-Datensätze sind das Ergebnis manueller Annotation und wurden zum

Teil seit mehreren Jahren nicht aktualisiert.

Zunächst lässt sich feststellen, dass die drei zu vergleichenden V-Gen-Datenbanken für alle Loci eine verschiedene Anzahl von V-Genen enthalten (Tabelle 2.5). Mit Ausnahme des humanen IgH-Locus enthält VBASE2 die meisten V-Gene.

Tabelle 2.5: Anzahl der V-Gene für die Immunglobulinloci von Mensch und Maus in den Datensätzen von VBASE2, Vbase, IMGT-Referenzsequenzen und Almagro

Mensch	IGHV	IGKV	IGLV
VBASE2	270	146	128
Vbase	349	112	118
IMGT	237	72	78

Maus	IGHV	IGKV	IGLV
VBASE2	379	201	5
ABG	288	82	-
IMGT	234	132	14

Vergleicht man die V-Gen-Sequenzen untereinander, wird deutlich, dass die V-Gen-Datensätze nur partiell identisch sind. So gibt es einerseits Sequenzen, die nicht in VBASE2, aber in den anderen Datenbanken vorhanden sind. Dies gilt am meisten für die IgH-V-Gene von Vbase: Von den 349 Vbase-Sequenzen sind 196 in VBASE2 nicht enthalten. Andererseits gibt es zahlreiche Sequenzen, die in VBASE2 vorhanden sind, in den anderen Datenbanken jedoch fehlen. Eine Untersuchung dieser unikalen VBASE2-V-Gene zeigt, dass es sich dabei überwiegend um Klasse-2-V-Gene handelt (Tabelle 2.6). Es gibt jedoch auch zahlreiche VBASE2-Klasse-1-V-Gene, die in anderen Datenbanken nicht enthalten sind. Der große Anteil muriner IgH-V-Sequenzen ist dabei durch die

aktuelle Sequenzierung des 129/Sv-Schwerekettenlocus zu erklären, die im zweiten Ergebnisteil dieser Arbeit ausgewertet wird. Die V-Gen-Allele dieses Locus waren bei der letzten Aktualisierung der ABG- und IMGT-Datensätze noch nicht veröffentlicht.

Tabelle 2.6: Klassifizierung unikaler VBASE2-V-Gene

Angegeben ist die Anzahl und Klassenzugehörigkeit der V-Gene, die nicht im Vbase-, IMGT-, beziehungsweise ABG-Datensatz enthalten sind.

Mensch	Datenbank	Klasse 1	Klasse 2	Klasse 3	Summe
IGHV	Vbase	1	109	6	116
	IMGT	5	170	6	181
IGKV	Vbase	4	45	5	54
	IMGT	6	86	5	97
IGLV	Vbase	5	42	5	52
	IMGT	5	63	5	73

Maus	Datenbank	Klasse 1	Klasse 2	Klasse 3	Summe
IGHV	ABG	79	206	15	300
	IMGT	74	206	14	294
IGKV	ABG	3	63	4	70
	IMGT	51	105	4	160
IGLV	ABG	-	-	-	-
	IMGT	0	0	0	0

2.1.2 Die Datenbank VBASE2

Die VBASE2-Datenbank ist im Internet unter http://www.vbase2.org zu erreichen. VBASE2 ist eine dynamische Datenbank und stellt die Keimbahn-V-Gen-Sequenzen zur Verfügung, die in dem oben beschriebenen automatischen Analyse-Prozess ermittelt wurden. Neben der Nukleotidsequenz und allgemeinen Informationen über das jeweilige V-Gen werden auch Verweise auf die Original-Sequenzen gespeichert, die dem V-Gen und dessen Zuordnung als Keimbahn-V-Gen zugrunde liegen. Ebenfalls werden die EMBL-IDs aller Rearrangements eines V-Gens aufgeführt. Weiterhin wird angezeigt, unter welcher ID ein V-Gen in der Kabat-, IMGT/LIGM- oder Vbase-Datenbank gespeichert ist. Die Verknüpfung mit der EMBL-Bank, Ensembl und anderen V-Gen-Datenbanken macht VBASE2 zu einer integrativen Plattform zwischen verschiedenen Informationssystemen.

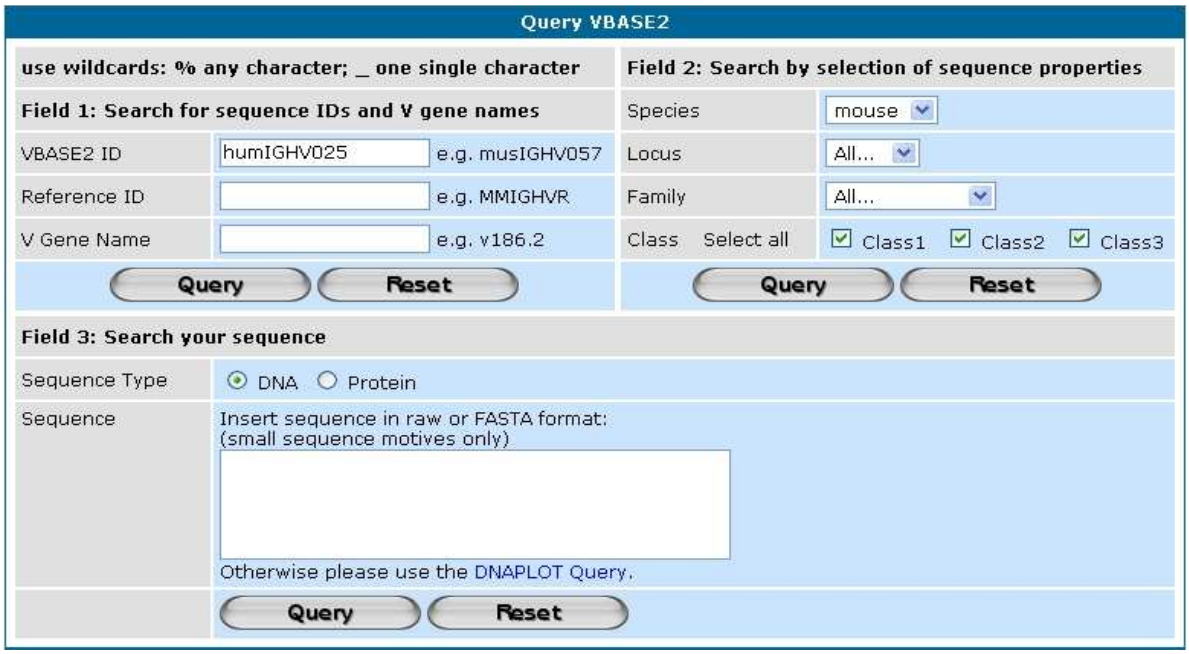
2.1.2.1 Abfrage von Daten über das Webinterface

Für einen Zugriff auf die Datenbank stehen zwei Möglichkeiten der Anfrage zur Verfügung: die Text-basierte Suche und die Sequenz-basierte Suche.

Bei der Text-basierten Suche können Suchbegriffe wie VBASE2-ID, V-Gen-Name und EMBL-ID eingegeben werden (Abbildung 2.7). Weiterhin kann die Suche durch Wahl der Spezies, des Locus und der V-Gen-Familie definiert werden. Drittens kann eine Protein- oder DNA-Sequenz für die Suche verwendet werden. Da bei der Text-basierten Suche aber nach einer exakten Zeichenfolge gesucht wird, werden hier nur solche Sequenzen gefunden, die eine mit der Anfrage identische Subsequenz aufweisen. Dieses Suchfeld ist daher geeignet für die Identifizierung von Sequenzen aufgrund von kurzen Sequenz-Motiven.

Abbildung 2.7: Die 'Direct Query'-Oberfläche für Text-basierte Suchanfragen

Die Suchmaske besteht aus drei alternativ zu gebrauchenden Feldern, in denen unterschiedliche Suchkriterien eingegeben werden können. Field 1 erlaubt die Suche nach V-Gen-IDs und -Namen, in Field 2 kann das Suchergebnis durch Auswahl von Spezies, Locus, Familie und VBASE2-Klasse ausgewählt werden, und Field 3 ermöglicht die Suche nach DNA- oder Proteinsequenzen.



Für die Suche einer vollständigen V-Gen-Sequenz in VBASE2 bietet die 'DNAPLOT Query'-Seite eine Ähnlichkeitssuche mit dem DNAPLOT-Programm. Bei Eingabe einer V-Gen-Sequenz wird das Alignment des V-Gens mit den ähnlichsten VBASE2-V-Genen ausgegeben. Über die VBASE2-ID der Sequenzen gelangt man zum V-Gen-Eintrag.

Ein VBASE2-V-Gen-Eintrag (Abbildung 2.8) zeigt Informationen über V-Gen-Namen, Familie und VBASE2-Klasse an. Das Feld 'Cross References' enthält die IDs der Originaleinträge der EMBL-Bank und Positionen in den Ensembl-Chromosomen, die dem Eintrag zu Grunde liegen. Handelt es sich bei der Original-Sequenz um einem BAC oder Contig, wird die Position des V-Gens

innerhalb der Sequenz angegeben. Die EMBL-IDs verweisen über SRS auf die EMBL-Bank, die Ensembl-Angabe verweist auf den ContigView des Ensembl Genom-Browsers in der entsprechenden Position. Im Feld 'Cross References' wird auf Einträge in den V-Gen-Datenbanken Vbase, IMGT/LIGM und Kabat verwiesen, sofern das V-Gen in der entsprechenden Datenbank vorhanden ist.

Abbildung 2.8: Beispiel eines VBASE2-V-Gen-Eintrags

Die Abbildung zeigt einen Klasse-1-Eintrag des humanen IgH-Locus (VBASE2-Version vom April 2005).

General Information	
VBASE2 ID	humIGHV025
Class	class 1: genomic and rearranged references
Date	2005-04-21
V Gene Name(s)	V3-15+, IGHV3-15*02
Family	IGHV3
Locus	IGHV
Species	human
Source References	
Genomic Sequence	EMBL: HSIGH315X
Rearranged Sequence	EMBL: AF209739, AY607526, AY582401, HSIGHVAI, AF297162, AY429997
Cross References	
VBASE (Search VBASE)	V3-15+
IMGT	HSZ80428, HSZ80374, HSIGH315X, HSB7G3B06, HSB7G3B05, BD013939, AY429875, AF077472 more...
KABAT (ftp KABAT)	KABID_044750, KABID_044669, KABID_044585, KABID_044584, KABID_021833, KABID_000552, KABID_000502
Features	
Protein Translation	1 EVQLVESGGA LVKPGGSLRL SCAASGFTFS NAWMSWVRQA PGKGLEWVGR 50 51 IKSKTDGGTT DYAAPVKGRF TISRDDSKNT LYLQMNSLKT EDTAVYYCTT

Nucleotide Sequence Structure						
	FR1	1..75	CDR1	76..99	1st_CYS	64..66
	FR2	100..150	CDR2	151..180	CONSERVED_TRP	106..108
	FR3	181..294	CDR3	295..>300	2nd_CYS	292..294

Nucleotide Sequence	
Length	300 bp
Sequence	1 GAGGTGCAGC TGGTGGAGTC TGGGGGAGCC TTGGTAAAGC CTGGGGGGTC 50 51 CCTTAGACTC TCCTGTGCAG CCTCTGGATT CACTTTCAGT AACGCCTGGA 100 101 TGAGCTGGGT CCGCCAGGCT CCAGGGAAGG GGCTGGAGTG GGTTGGCCGT 150 151 ATTAAAAGCA AAACTGATGG TGGGACAACA GACTACGCTG CACCCGTGAA 200 201 AGGCAGATTC ACCATCTCAA GAGATGATTC AAAAAACACG CTGTATCTGC 250 251 AAATGAACAG CCTGAAAACC GAGGACACAG CCGTGTATTA CTGTACCACA

2.1.2.2 Integration von VBASE2 in bestehende Informationssysteme und Datenbanken

Die immunologischen Datenbanken des IMGT, die Kabat-Datenbank und die Vbase-Datenbank sind unabhängige Informationssysteme, die unterschiedliche Strategien verfolgen und für den Nutzer jeweils ihre eigenen Vorzüge haben. VBASE2 hat als neu geschaffene Datenbank das Ziel, die bestehenden Systeme miteinander zu verbinden. Da VBASE2 aus der EMBL-Bank und aus den Ensembl-Sequenzen generiert wird, stellt die Datenbank eine Plattform zwischen den primären Datenbanken einerseits und den immunologischen Datenbanken andererseits dar.

VBASE2 verweist auf mehrere tausend Einträge in der EMBL-Bank, die als Referenz-Sequenzen für die V-Gen-Einträge ermittelt wurden. Umgekehrt verweist die EMBL-Bank auch auf VBASE2: in jeder der von VBASE2 ausgewählten Referenzen wird die VBASE2-ID unter 'Database Cross references' im EMBL-Bank-Eintrag angegeben. Im EMBL release 83 vom Juni 2005 verweisen 4605 Einträge auf einen VBASE2-Eintrag.

VBASE2-V-Gene, deren Sequenzen in den Chromosomen des Ensembl-Assemblys vorkommen, verweisen auf die entsprechende Position im Ensembl Genom-Browser. Durch die Einrichtung eines DAS-Servers können die V-Gene von VBASE2 auch im Ensembl Contigview angezeigt werden (Abbildung 2.9). Dazu muss der Browser mit dem VBASE2-DAS-Server unter http://www.dnaplot.com/das verbunden werden. Die Datensätze 'vbase2_mouse' und 'vbase2_human' können dabei einzeln ausgewählt werden. Im Contigview werden die V-Gene in FRs und CDRs unterteilt und unter Angabe der konservierten Aminosäuren Cystein und Tryptophan dargestellt (Abbildung 2.9 B). Von den VBASE2-V-Genen in Ensembl gelangt man über ein

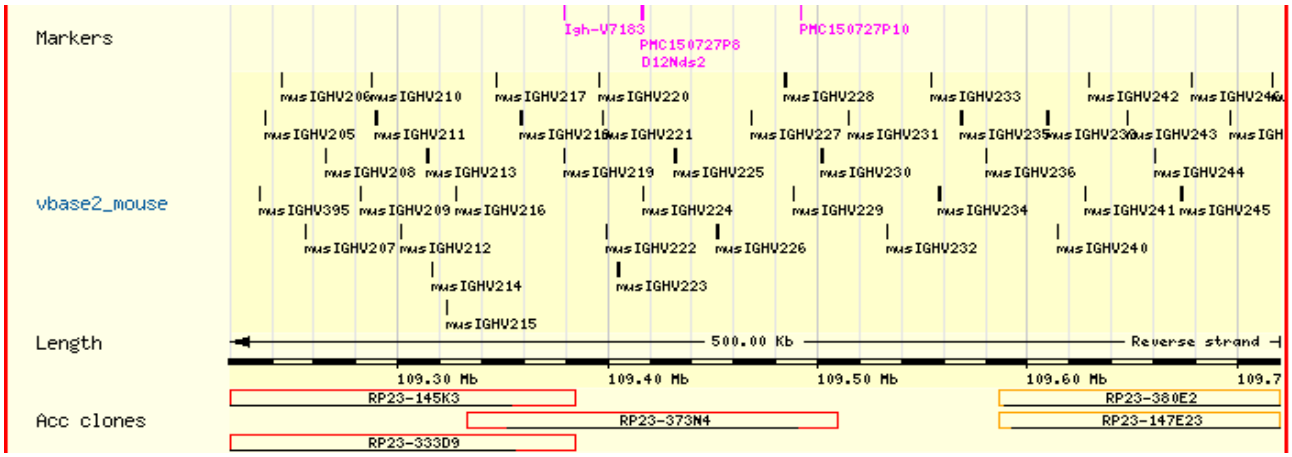
Kontextmenü wiederum zu den VBASE2-Datenbank-Einträgen.

VBASE2 verweist ebenfalls auf die untereinander autarken V-Gen-Datenbanken IMGT/LIGM, Kabat und, bei humanen V-Genen, auf Vbase. IMGT/LIGM enthält alle EMBL-Bank-Sequenzen, die als Immunglobuline annotiert sind, und liefert zusätzliche Informationen in sehr unterschiedlichem Umfang. Vbase enthält humane Keimbahn-V-Gen-Sequenzen, die durch einmalige manuelle Annotation aller V-Gene der EMBL-Bank ermittelt wurden. Die Kabat-Datenbank ist als erste V-Gen-Datenbank von großer historischer Bedeutung und ist unabhängig von EMBL-Bank/GenBank/DDBJ. Da alle drei Datenbanken von vielen Immunologen genutzt wurden und werden, ist die Angabe der in diesen Datenbanken verwendetem V-Gen-IDs oder -Namen ein Gewinn für die Nutzer von VBASE2. Ein ähnlicher Zusammenhang besteht bei der V-Gen-Benennung: Da in der Literatur verschiedene V-Gen-Nomenklaturen nebeneinander verwendet werden, werden in VBASE2 alle bekannten Namen eines V-Gens aufgeführt (siehe Abbildung 2.8).

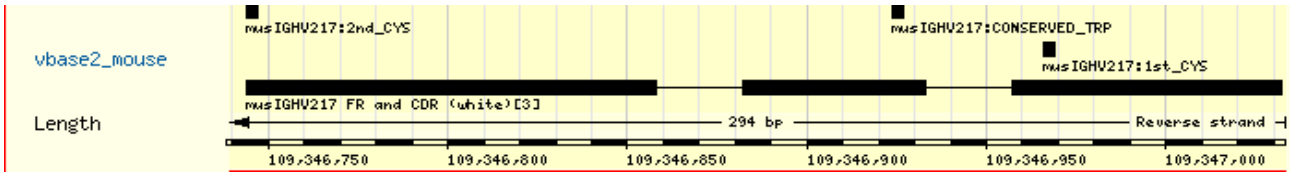
Abbildung 2.9: Anzeige von VBASE2-V-Genen im Ensembl Genom-Browser mit Hilfe des DAS-Servers

Abgebildet sind zwei Ausschnitte verschiedener Vergrößerungen des Maus-Chromosoms 12.

A: Übersicht über die V-Gene in einem 0,5 Mb Ausschnitt des murinen IgH-Locus.



B: Bereich eines V-Gens mit Darstellung der FRs (schwarze Balken) und konservierten Aminosäuren Cystein 1, Cystein 2 und Tryptophan.



2.1.3 Anwendung der automatischen V-Gen-Analyse: Immunglobuline in UniProtKB/TrEMBL

Im Rahmen eines Marie-Curie-Stipendiums wurde ein drei Monate dauerndes Praktikum am Europäischen Bioinformatik-Institut (EBI) durchgeführt. In diesem Zeitraum wurde auf der Basis des automatischen V-Gen-Analyse-Prozesses und des VBASE2-Datensatzes eine Anwendung entwickelt, die den kontrollierten Eingang von V-Genen in UniProtKB/TrEMBL erlaubt. TrEMBL enthält die kodierenden Sequenzen der EMBL-Bank als Aminosäure-Sequenzen und repräsentiert damit den automatisch erzeugten Teil der UniProt Knowledgebase (UniProtKB) [Bairoch et al., 2005]. Immunglobulin-Sequenzen werden bisher von TrEMBL ausgeschlossen, weil die Vielzahl an ähnlichen und redundanten Sequenzen den Datensatz überfluten würde [Boeckmann et al., 2003]. Für diesen Zweck wird ein Text-basierter Filter verwendet, der den Eingang von Immunglobulin-Sequenzen durch die Identifizierung Immunglobulin-typischer Annotation verhindern soll. Eine Prüfung der TrEMBL-Sequenzen zeigte jedoch schnell, dass der derzeit verwendetete Filter nicht ausreichend ist. Es gibt zahlreiche Immunglobulin-TrEMBL-Einträge, die als Sequenzen ohne Annotation durch den Text-Filter nicht erkannt werden. Viele davon stammen aus cDNA-Sequenzierprojekten [Strausberg et al., 2002; Okazaki et al., 2002]. Die erste Aufgabe bestand deshalb darin, einen Sequenz-basierten Filter für V-Gene in TrEMBL zu erstellen. Die zweite Aufgabe war die Erstellung einer Liste von EMBL-Bank-Einträgen, die für die Erzeugung eines umfassenden, aber nicht-redundanten V-Gen-Datensatzes in TrEMBL genutzt werden kann. Da die TrEMBL-Datenbank automatisch erzeugt wird, sollten beide Prozesse so weit wie möglich automatisiert werden.

2.1.3.1 Sequenzbasierter Filter für Immunglobuline in UniProtKB/-TrEMBL

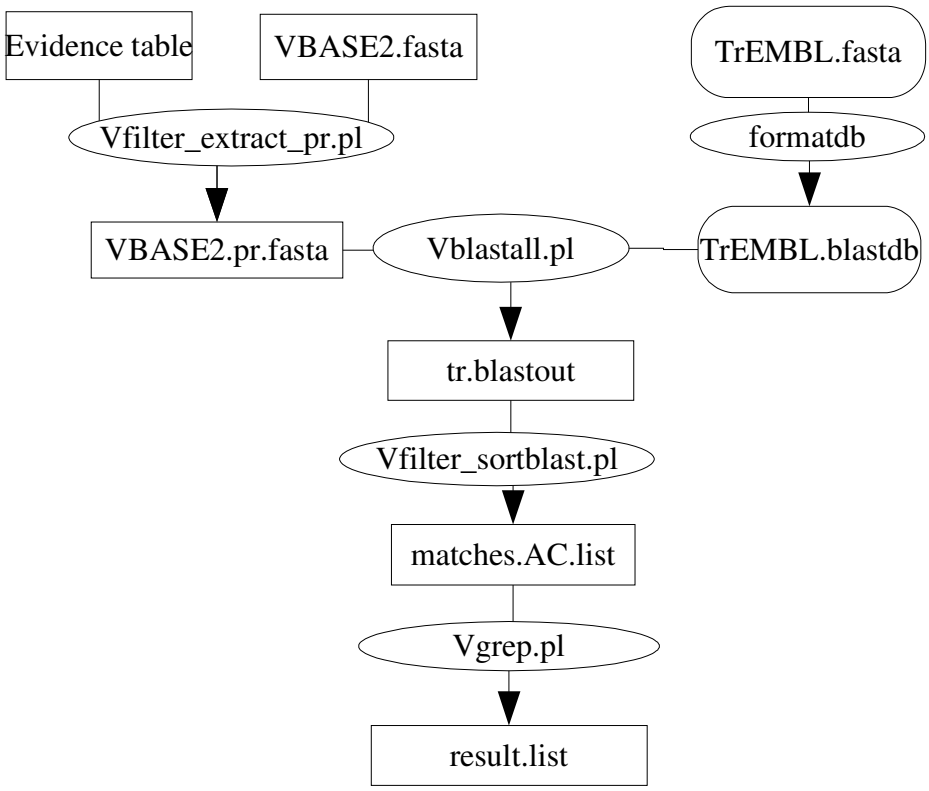
Der Programm-Ablauf des Filters für V-Gen-Sequenzen ist in Abbildung 2.10 und Tabelle 2.7 beschrieben. Die verwendeten Perl-Skripte sind teilweise an die Skripten des V-Gen-Analyse-Prozesses angelehnt. Für die Erkennung von V-Gen-Sequenzen auf Aminosäureebene werden die VBASE2-Klasse-1-DNA-Sequenzen zunächst in Aminosäure-Code übersetzt und auf die Region FR1 bis FR3 begrenzt. Diese Peptidsequenzen von etwa 97 Aminosäuren Länge dienen als Input für eine BLAST-Suche [Altschul et al., 1997] in TrEMBL. Das BLAST-Ergebnis wird auf minimale Alignment-Länge und -Identität gefiltert. Sequenzen mit einer Alignment-Länge von mindestens 80 Aminosäuren und einer Identität von mindestens 80% werden automatisch als Immunglobulin-Sequenzen identifiziert.

Tabelle 2.7: Daten und Programme des Filters für Immunglobulin-Aminosäure-Sequenzen

Programm	Input	Funktion
Vfilter_extract_pr.pl	VBASE2.nt.fasta, evidence.table	Übersetzt die VBASE2-Klasse-1-Sequenzen aus der Datei 'evidence.table' in Aminosäure-Sequenzen bis zum Ende von FR3.
formatdb	TrEMBL.fasta	Erstellt eine BLAST-Datenbank für das NCBI-blastall-Programm.
Vblastall.pl	VBASE2.pr.fasta	Startet das blastall-Programm zur Suche der VBASE2-Aminosäure-Sequenzen in TrEMBL.
Vfilter_sortblast.pl	Tr blastout	Filtert das BLAST-Ergebnis nach Treffern definierter Alignment-Länge und -Identität.
Vgrep.pl	matches.AC.list	Entfernt die Immunglobulin-Sequenzen, die in TrEMBL bewusst eingepflegt wurden, aus dem Filter-Ergebnis.

Abbildung 2.10: Daten und Programme des Filters für Immunglobulin-Aminosäure-Sequenzen

Der Filter durchsucht die TrEMBL-Datenbank nach Sequenzen von V-Regionen. Daten sind in Rechtecken, Programme in Elipsen und Datenbanken in abgerundeten Rechtecken dargestellt. Erläuterungen zu Daten und Programmen sind in Tabelle 2.7 angegeben.



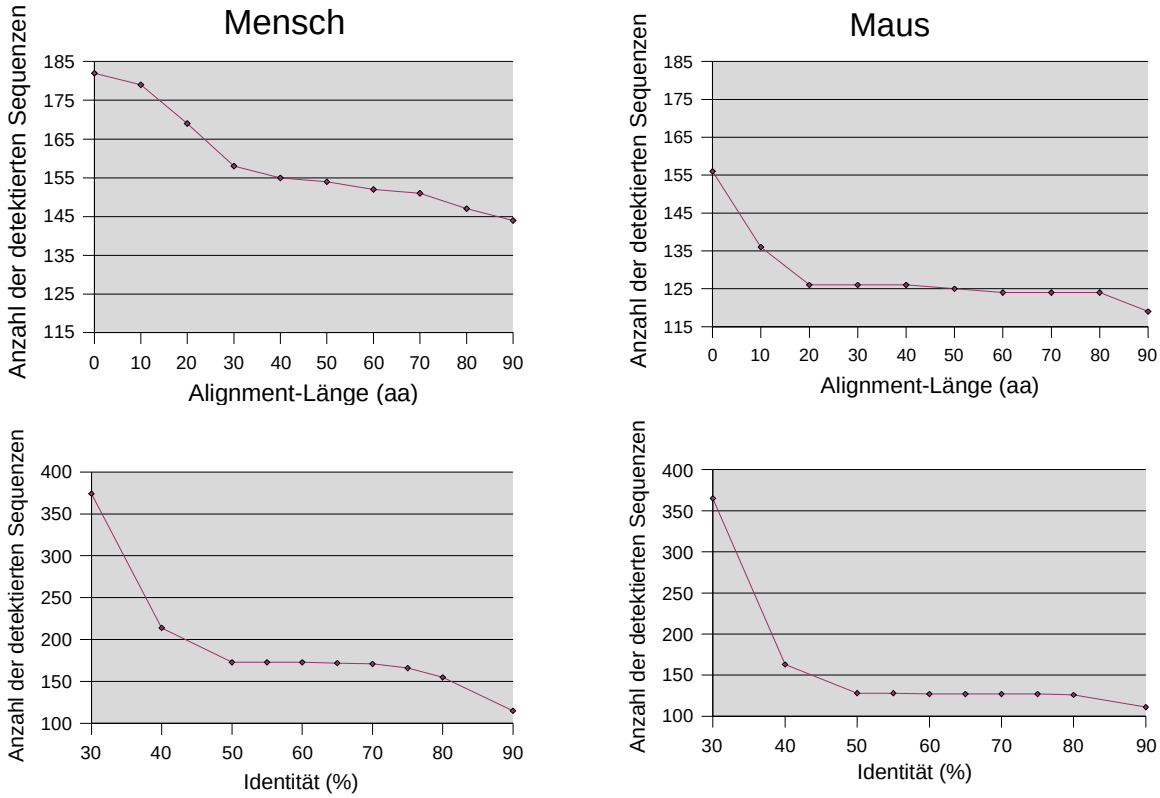
Um die Parameter Alignment-Länge und -Identität zu untersuchen, wurde jeweils einer der Parameter bei konstantem anderen Parameter variiert (Abbildung 2.11). Das Ergebnis zeigt, dass ab einem bestimmten Grenzwert die Anzahl der detektierten Sequenzen deutlich steigt. Beim Parameter Identität liegt der Grenzwert bei fünfzig Prozent. Bei der Alignment-Länge liegt der Grenzwert für murine Sequenzen bei zwanzig Aminosäuren, die Kurve der humanen Sequenzen verläuft weniger deutlich. Den Kurven ist weiterhin zu entnehmen, dass es nur wenige Sequenzen gibt, die den Immunglobulin-Sequenzen 'ein bisschen' ähnlich sind, da die Steigung im mittleren Bereich der Kurve sehr gering ist. Dies bestätigt auch die unter 2.1.1.3.1 angestellten

Untersuchungen zum Einfluss der BLAST-Input-Sequenzen auf das BLAST-Ergebnis in der EMBL-Bank. Im Fall der BLAST-Suche in TrEMBL ergab die manuelle Analyse der für positiv befundenen Sequenzen, dass die meisten der im mittleren Kurvenbereich detektierten Sequenzen tatsächlich Immunglobuline sind. Einige ließen sich jedoch weder durch Angaben zur Art und Herkunft der Sequenz noch durch Alignment mit dem DNAPLOT Webservice eindeutig als Immunglobuline identifizieren. Diese möglicherweise falsch-positiven Sequenzen dürfen dem TrEMBL-Datensatz nicht entzogen werden. Deshalb wurde die Grenze zur automatischen V-Gen-Detektion bei achtzig Prozent Identität und achtzig Aminosäuren Länge belassen. Sequenzen mit einer Ähnlichkeit von fünfzig bis achtzig und einer Alignment-Länge von vierzig bis achtzig Aminosäuren werden dagegen als mögliche Immunglobuline ('ambiguous') eingestuft. Durch manuelle Prüfung, unter anderem durch Alignment mit dem DNAPLOT-Webservices, können diese Sequenzen als Positive oder Negative eingeordnet werden.

Insgesamt wurden im TrEMBL-Datensatz vom September 2004 170 humane und 134 murine V-Gen-Sequenzen detektiert. Die EMBL-Zugangsnummern (ACs) der Sequenzen sind in Anhang I.1A aufgeführt.

Abbildung 2.11: Filter-Parameter Alignment-Länge und -Identität

Im Filter-Prozess werden V-Gene in TrEMBL durch Vergleich mit VBASE2-V-Genen identifiziert. Für die Erkennung sind dabei die Länge und die prozentuale Identität des Alignments ausschlaggebend. Dargestellt ist die Anzahl der detektierten Immunglobuline bei variabler Alignment-Länge und minimaler Sequenz-Identität.



2.1.3.2 Prozess zur Auswahl von Immunglobulin-Sequenzen für UniProtKB/TrEMBL

Die TrEMBL-Datenbank enthält einen nicht-redundanten Datensatz von vollständigen kodierenden Sequenzen der EMBL-Bank, der automatisch aus der CDS-Annotation des EMBL-Eintrags erzeugt wird [Bairoch et al., 2005]. Die Auswahl von Immunglobulin-Rearrangements für TrEMBL muss dementsprechend folgende Kriterien erfüllen:

1. Es liegt ein EMBL-Bank-Eintrag mit CDS-Annotation vor.

2. Die Rearrangements des TrEMBL-Immunglobulin-Datensatzes sollen nicht-redunante Keimbahn-V-Gene enthalten.
3. Stehen für ein Keimbahn-V-Gen in der EMBL-Bank mehrere Rearrangements zur Verfügung, so wird die erste Veröffentlichung der längsten Sequenz ausgewählt.

Der V-Gen-Selektionsprozess wurde so weit wie möglich dem Filter-Prozess analog gestaltet, so dass für beide Prozesse ähnliche oder die gleichen Skripte benutzt werden können. Der Ablauf des Selektionsprozesses ist in Tabelle 2.8 und Abbildung 2.12 dargestellt. Der Prozess beginnt, ebenso wie der Filter-Prozess, mit einer BLAST-Suche der VBASE2-Klasse-1-Sequenzen auf Proteinebene. In diesem Fall werden SWISS-PROT und TrEMBL durchsucht, um festzustellen, welche der VBASE2-Sequenzen bereits in der UniprotKB vorhanden sind. Neue V-Gen-Sequenzen werden ausgegeben, und mit Hilfe der VBASE2-Referenzen ('evidence.table') wird nach den genannten Kriterien ein EMBL-Bank-Eintrag ausgewählt. Das Kriterium der Vollständigkeit kann im Fall der Immunglobuline nicht eingehalten werden, da die Sequenz der konstanten Region unabhängig von den V-Gen-Sequenzen und Rearrangements aufgeklärt wurde [Maus: Shimizu et al., 1982; Mensch: Hofker et al., 1989]. Sequenzierungen, die die variable Region des Antikörpers aufklären, brechen deshalb gewöhnlich vor Beginn der konstanten Region ab. Insgesamt wurden nur vier humane und zwei murine vollständige CDS von Keimbahn-V-Gen-Sequenzen in der EMBL-Bank gefunden, beide kodieren leichte Ketten, deren konstante Region nur etwa ein Drittel so groß ist wie die konstante Region der schweren Kette.

Durch den Selektionsprozess wurden 120 humane und 158 murine EMBL-Bank-

Einträge ausgewählt und als neue Immunglobulin-Sequenzen für TrEMBL vorgeschlagen (Anhang I.1B).

Der Filter wurde für murine und humane V-Gen-Sequenzen entwickelt, ist jedoch prinzipiell auf alle Spezies, deren Immunglobulinloci eine vergleichbare Struktur haben, und auf T-Zell-Rezeptor-Loci anwendbar. Darüber hinaus ist es denkbar, den sequenzbasierten Filter auch für andere Arten von Sequenzen anzuwenden, die in TrEMBL nicht erwünscht sind. Ein naheliegendes Beispiel sind die Sequenzen der konstanten Regionen der Immunglobuline. Für eine erweiterte Anwendung des Filters, insbesondere für die Anwendung auf Nicht-V-Gen-Sequenzen, ist eine sorgfältige Prüfung der Filter-Parameter erforderlich.

Der Prozess soll nach Aussage von Maria Jesus Martin, Koordinatorin von UniProtKB/TrEMBL am EBI, Ende 2005 in den TrEMBL Generationsprozess integriert und die ausgewählten Immunglobuline in den Datensatz eingepflegt werden.

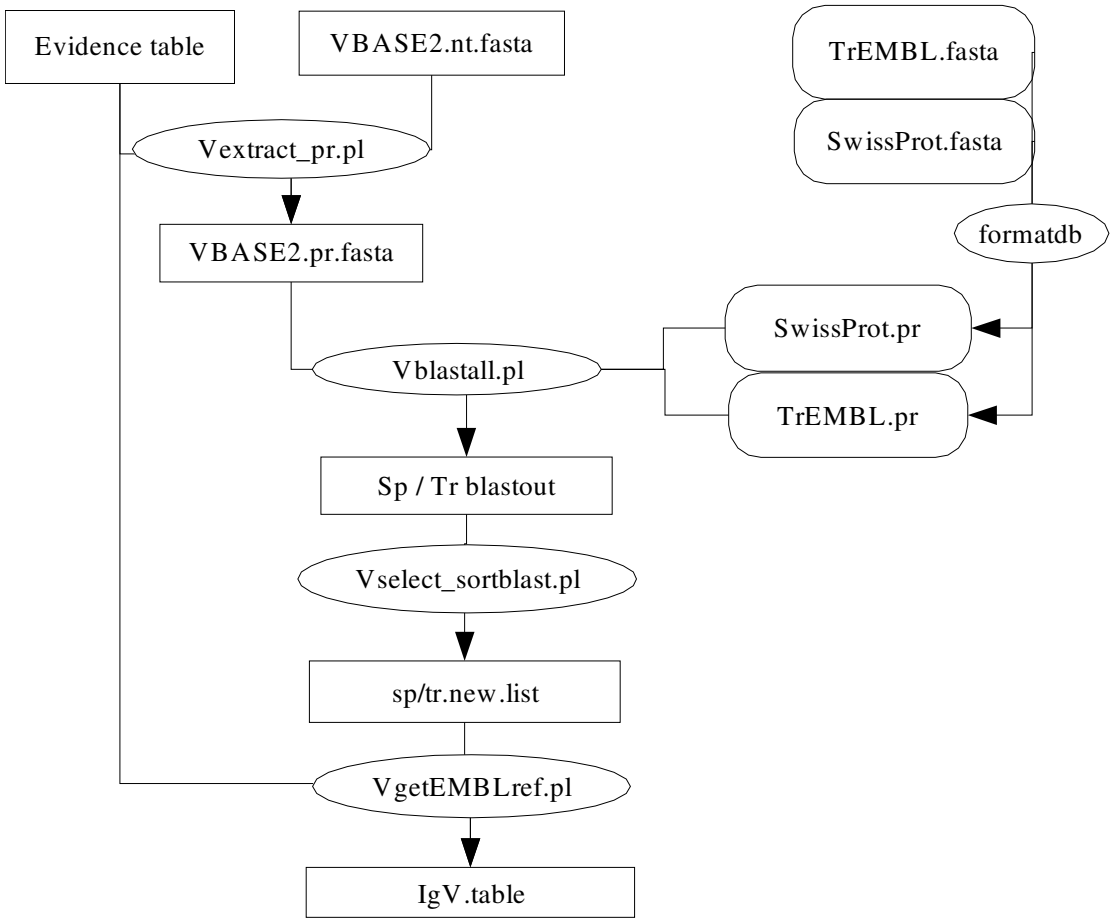
Tabelle 2.8: Daten und Programme zur Selektion von Immunglobulin-Sequenzen für TrEMBL

Programm	Input	Funktion
Vextract_pr.pl	VBASE2.nt.fasta evidence.table	Übersetzt die VBASE2-Klasse-1-Sequenzen aus der Datei 'evidence.table' in Aminosäure-Sequenzen bis zum Ende von FR3.
formatdb	TrEMBL.fasta, SwissProt.fasta	Erstellt BLAST-Datenbanken von SWISS-PROT und TrEMBL für das NCBI-blastall-Programm.
Vblastall.pl	VBASE2.pr.fasta	Startet das blastall-Programm zur Suche der VBASE2-Aminosäure-Sequenzen in SWISS-PROT und TrEMBL.
Vselect_sortblast .pl	Sp / Tr blastout	Filtert das BLAST-Ergebnis in der Datei 'blastout' nach Treffern definierter Alignment-Länge und -Identität.

Programm	Input	Funktion
VgetEMBLref.pl	sp/tr.new.list	Wählt aus den EMBL-Bank-Einträgen eines V-Gens mit CDS-Annotation den ältesten Eintrag der längsten Sequenz aus.

Abbildung 2.12: Daten und Programme zur Selektion von Immunglobulin-Sequenzen für UniProtKB/TrEMBL

Der V-Region-Selektionsprozess durchsucht zunächst UniProtKB mit einer BLAST-Suche der aktuellen VBASE2-Sequenzen nach bereits vorhandenen V-Regionen und schließt diese von der Analyse aus. Für neue Sequenzen wird nach den im Text genannten Kriterien ein EMBL-Bank-Eintrag ausgewählt. Daten sind in Rechtecken, Programme in Elipsen und Datenbanken in abgerundeten Rechtecken dargestellt.



2.2 *In-silico*-Charakterisierung des Immunglobulin-Schwerekettenlocus (IgH-Locus) des Mausstammes 129/Sv

2.2.1 Sequenz-Assemblierung

Die Sequenzierung des Immunglobulin-Schwerekettenlocus (IgH-Locus) des Mausstammes 129/Sv basiert auf einer physikalischen Karte, die in der Arbeitsgruppe von Roy Riblet am Torrey Pines Institute for Molecular Studies durch Verwendung von bakteriellen artifiziellen Chromosomem (BACs) angefertigt wurde. Eine mit Hilfe der Karte erstellte Abfolge überlappender BACs wurde in der Abteilung Genomanalyse der Gesellschaft für Biotechnologische Forschung sequenziert und mit dem Programm Gap4 assembliert, so dass eine Sequenz aus 32 Contigs mit 1555244 bp Länge entstand (Tabelle 2.9). Die physikalische Karte konnte dabei im Wesentlichen bestätigt werden. Es soll hier bemerkt werden, dass die Sequenzierung und Assemblierung aufgrund ausgeprägter interner Sequenzähnlichkeiten und repetitiver Bereiche deutlich anspruchsvoller war als bei durchschnittlichen Sequenzen. Die mit Gap4 assemblierte Sequenz ist die Grundlage der hier beschriebenen Analyse.

Zur Fortsetzung des Assemblierungsprozesses und Validierung der Assemblierung wurden BLAST-Suchen, MultiPipMaker-Analysen und die Ensembl-Webseite benutzt. Mit Hilfe der genannten Anwendungen konnten sieben der 31 Lücken geschlossen werden. Das Ergebnis der Assemblierung (Tabelle 2.10) umfasst 25 Contigs, wobei dreizehn davon zum nicht fertiggestellten BAC 4k8 gehören und deshalb in Tabelle 2.10 nicht dargestellt sind. Die Contigs C02 bis C14 umfassen eine Sequenz von 1382053 bp.

Neunzehn der 21 in Tabelle 2.10 aufgeführten BACs wurden bei der EMBL-Bank eingereicht, mit Ausnahme des 4k8-BACs sind sie öffentlich zugänglich. Nachträglich wurde auch die assemblierte Sequenz der Contigs C02 bis C14 eingereicht, die nun unter der EMBL-AC AJ851868 verzeichnet ist. Die EMBL-ACs AJ851869-AJ851885 verweisen auf die Gesamtsequenz AJ851868.

Tabelle 2.9: Status der Sequenzierung zu Beginn der Analyse

Die Reihenfolge der Klone entspricht ihrer Anordnung auf dem Chromosom. Der EMBL-Bank-Eintrag AJ972404 ist noch nicht öffentlich zugänglich. Die Sequenzierung der Klone 436c3 und 456e6 wurde eingestellt, weil die verbleibenden Lücken mit den angewandten Sequenziermethoden nicht geschlossen werden können. Die Klone 462f18 und 137c2 wurden zur Unterstützung der Assemblierung der überlappenden Klone teilweise sequenziert, jedoch nicht als eigenständige Klone bei der EMBL-Bank eingereicht.

Klon-Name	Status	Anzahl der Contigs	Größe [bp]	EMBL-AC
4k8	Assemblierung	14	154326	AJ972404
260d2	abgeschlossen	1	111841	AJ851875
197b5	abgeschlossen	1	119600	AJ851873
396l2	abgeschlossen	1	134884	AJ851879
462f18	Assemblierung	7	238132	-
356c12	Assemblierung	5	129618	AJ851877
197k2	Assemblierung	2	129052	AJ851874
363p17	abgeschlossen	1	122898	AJ851878
473a9	abgeschlossen	1	144659	AJ851882
75m10	abgeschlossen	1	122986	AJ851871
511o23	abgeschlossen	1	95647	AJ851883
436c3	abgeschlossen	2	151635	AJ851880
456e6	abgeschlossen	2	103672	AJ851881
189g9	Assemblierung	4	123974	AJ851872
355e24	abgeschlossen	1	105638	AJ851876
62g14	Assemblierung	2	114728	AJ851870.1
150m14	abgeschlossen	1	115366	AJ851884

Klon-Name	Status	Anzahl der Contigs	Größe [bp]	EMBL-AC
137c2	Assemblierung	75	232073	-
167c1	abgeschlossen	1	113205	AJ851885
363d14	Assemblierung	3	183143	AJ851868
34h6	Assemblierung	2	178.495	AJ851869

Tabelle 2.10: Ergebnis der Assemblierung

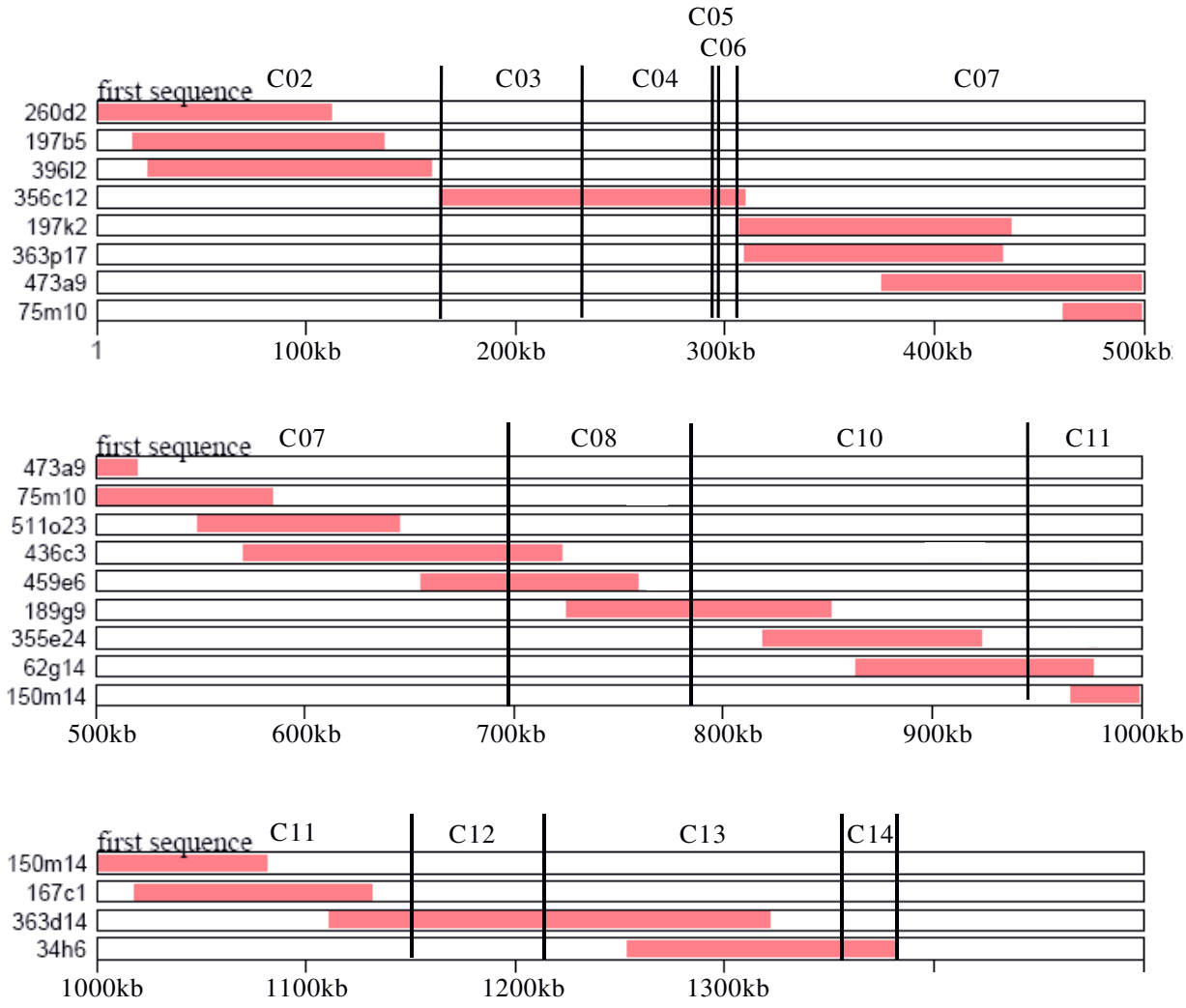
Die Reihenfolge der Contigs C04, C05 und C06 wurde mit Hilfe der homologen Sequenz vom Mausstamm C57BL/6 bestimmt und ist daher als hypothetisch anzusehen.

Contig-Name	Länge (bp)	BACs und -Fragmente
C02	163966	260d2, 197b5, 396l2, 462f18
C03	66585	462f18, 356c12
C04	61095	356c12
C05	2898	356c12
C06	9168	356c12
C07	392389	356c12, 197b5, 363p17, 473a9, 75m10, 511o23, 436c3, 456e6
C08	87340	436c3, 456e6, 189g9
C10	158560	189g9, 355e24, 62g14
C11	205702	62g14, 150m14, 167c1, 363d14
C12	64326	363d14
C13	140039	363d14, 34h6
C14	28885	34h6

Mit dem Programm PipMaker wurde aus der assemblierten Sequenz und den BAC-Sequenzen eine Karte erstellt, aus der die Positionen der BACs und der Contigs hervorgehen (Abbildung 2.13). Der Klon 4k8 liegt JH-distal vom Klon 260d2, hat jedoch keine Überlappung mit 260d2. Deshalb wurde er auf der Karte nicht verzeichnet.

Abbildung 2.13: BAC-Karte des IgH-Locus

Die Abbildung zeigt die Überlappung der BACs als Pipmaker-Output. Senkrechte Linien kennzeichnen die Positionen der Lücken zwischen den Contigs. Zur Erstellung des Pips wurde der 1,38 Mb große Supercontig in 500 kb-Fragmente geteilt, wobei das 380 kb-Fragment mit 120 kb 'ACTG'-Sequenz auf 500 kb verlängert wurde. Diese Supercontig-Fragmente wurden als 'first sequence' für drei BLASTZ-Alignments gegen die jeweiligen BACs mit dem Multipipmaker verwendet. Die Positionen der Contigs wurden aus weiteren Pipmaker-Karten in diese Abbildung übertragen. Ein Contig des BACs 462f18 deckt das Ende von C02 ab, der BAC wurde jedoch nicht separat assembliert und ist deshalb in dieser Karte nicht dargestellt.



2.2.2 Repetitive Elemente und interne homologe Bereiche

2.2.2.1 Repetitive Elemente

Die im Zusammenhang mit der Veröffentlichung des Mausgenoms 2002 angestellte erste Untersuchung der gesamten genomischen Sequenz der Maus ergab, dass etwa 39% des Mausgenoms aus sogenannten 'interspersed repeats' besteht, also aus DNA-Elementen, die in unterschiedlichen Kopienzahlen über das ganze Genom verteilt vorkommen [Waterston et al., 2002]. Den größten Anteil, insgesamt 37% des Mausgenoms, nehmen dabei Retrotransposons ein. Diese mobilen genetischen Elemente können transkribiert, mit Hilfe einer reversen Transkriptase in cDNA übersetzt und wieder in die genomische DNA integriert werden [Review: Kazazian, 2004]. Man unterscheidet zwischen den virus-ähnlichen LTR-Retrotransposons und Non-LTR-Retrotransposons. Bei letzteren unterscheidet man wiederum zwischen den LINE-Elementen (long interspersed nucleotide elements) und den SINE-Elementen (short interspersed nucleotide elements). LINEs kodieren, wie LTR-Retrotransposons, eine eigene reverse Transkriptase und werden deshalb auch als autonome Retrotransposons klassifiziert. SINEs benötigen für die Verbreitung eine reverse Transkriptase von einem aktiven LINE-Element [Dewannieux et al., 2003; Dewannieux und Heidmann, 2005]. Die Untersuchungen des Mouse Genome Sequencing Consortium (MGSC) ergaben einen durchschnittlichen Anteil von 10% LTR Retrotransposons, 19% LINEs und 8% SINEs im Genom der Maus. Der durchschnittliche GC-Gehalt der DNA wurde mit 42% angegeben.

Zur Detektion von repetitiver DNA im IgH-Locus wurde die Sequenz der Contigs C02 bis C14 mit dem Repeatmasker-Programm untersucht. Das Ergebnis (Tabelle 2.12) zeigt einen Gesamtgehalt an repetitiver DNA von 53% bei einem

GC-Level von 42%. Der GC-Gehalt des proximalen Teils des IgH-Locus entspricht damit exakt dem durchschnittlichen GC-Gehalt des Maus-Genoms, der Anteil an repetitiven Elementen ist aber deutlich erhöht. Dies betrifft insbesondere den Anteil an LINE1-Elementen, der bei 34% liegt. Auch die Anzahl der LTR-Retrotransposon ist im IgH-Locus um ein Drittel höher als im Gesamtgenom. Dagegen liegt der Anteil an SINEs deutlich unter dem Durchschnitt.

Eine getrennte Betrachtung von konstanter und variabler Region zeigt, dass der überdurchschnittliche Anteil an LINE-Sequenzen in der variablen Region besonders ausgeprägt ist: In der konstanten Region werden nur 19% der Sequenz als LINE1 identifiziert. Der Anteil an LTR-Elementen ist mit 6% in der konstanten Region sogar unterdurchschnittlich.

Tabelle 2.12: Gehalt an repetitiver DNA in 1,38 Mb des IgH-Locus von 129/Sv

Klassifizierung	Contigs C02-C14 (1,38 Mb)		C-Region (174 kb)
	Anzahl	Anteil [%]	Anteil [%]
SINEs	79	0,81	0,73
LINEs	507	33,69	19,16
LTR-Retrotransposons	321	15,16	6,33
DNA-Transposons	2	0,02	0,09
nicht klassifizierte Elemente	6	0,11	0
einfache Wiederholungen	394	2,32	6,98
Regionen geringer Komplexität	141	0,50	0,19

Innerhalb der konstanten Region gibt es repetitive Bereiche, die für den Immunglobulinlocus funktionelle Bedeutung haben: stromaufwärts der Exon-Gruppen für einzelnen Schwereketten-Isotpyen liegen die S-Regionen [Kataoka et al., 1980; Davies et al., 1980; Takahashi et al., 1980; Nikaido et al., 1981;

Kataoka et al., 1981]. Diese werden für den Klassenwechsel des Antikörpers benötigt und bestehen aus repetitiven Sequenzen unterschiedlicher Art und Länge. Vom Repeatmasker werden die insgesamt sieben S-Regionen nur teilweise erkannt (siehe Abbildung 2.15). Die Struktur der S-Regionen wird im Zusammenhang der konstanten Region im Abschnitt 2.2.3 besprochen.

2.2.2.2 Interne homologe Bereiche

Die multiplen Gensegmente der Immunglobulinloci sind vermutlich durch Duplikation und anschließende Diversifikation einiger Vorläufergene entstanden [Ota und Nei, 1994; Sitnikova und Su, 1998]. Um den vorliegenden Sequenzabschnitt des murinen IgH-Locus auf interne Sequenzwiederholungen zu untersuchen, wurde die Sequenz durch Erstellung eines Dotplots mit dem Programm Advanced PipMaker mit sich selbst verglichen. Das Programm erlaubt die farbige Markierung von Sequenzabschnitten, so dass die Positionen der V-, D- und J-Segmente und der Exone der konstanten Region im Dotplot dargestellt werden können. Das Ergebnis in Abbildung 2.14 zeigt, dass es in der V-, D- und C-Region des Locus zahlreiche interne homologe Bereiche gibt, die sich innerhalb der jeweiligen Region einmal oder mehrfach wiederholen. Fast sämtliche kodierenden Bereiche liegen innerhalb solcher Ähnlichkeitsbereiche. Die Anzahl und Länge der Wiederholungen ist bei der V-, D- und C-Region sehr unterschiedlich.

Repetitive Elemente, die vom Repeatmasker erkannt wurden (siehe 2.2.2.1), wurden vom Advanced PipMaker maskiert und sind deshalb im Dotplot nicht dargestellt. Im ebenfalls erzeugten Pip (Percent Identity Plot) sind Art und Positionen der maskierten Elemente jedoch eingezeichnet. Ein vollständiger Pip der Region ist in Anhang I.3 einzusehen. Ähnlich wie der Dotplot visualisiert

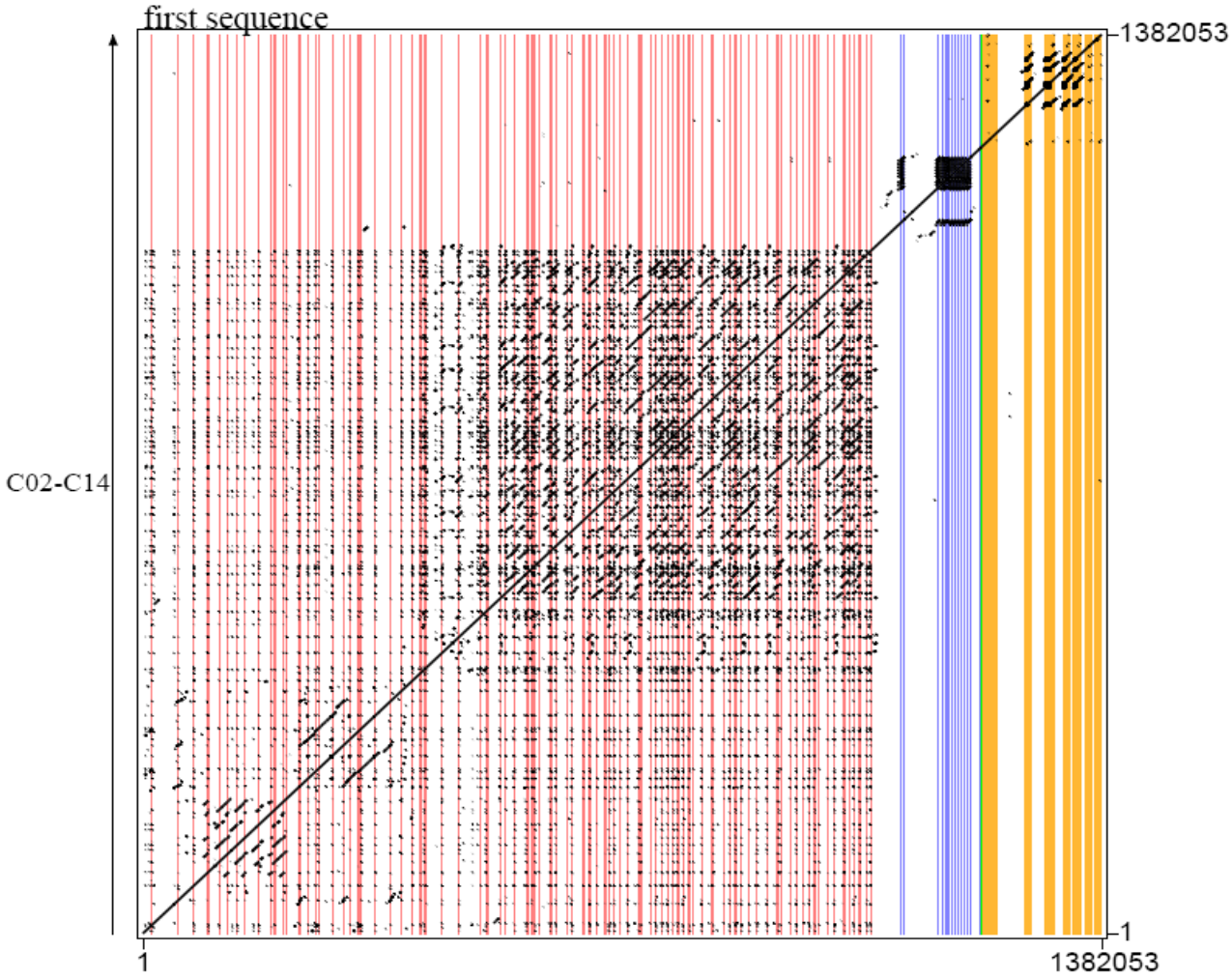
der Pip Sequenzwiederholungen, diese werden aber in höherer Auflösung und in Verbindung mit den vom Repeatmasker-Programm detektierten repetitiven Elementen dargestellt.

Innerhalb der V-Region gibt es zahlreiche kurze ähnliche Bereiche, die die V-Segmente abdecken und sich sehr oft wiederholen. Dabei beinhaltet der JH-proximale Teil der variablen Region einen deutlich abzugrenzenden Bereich mit besonders vielen homologen Sequenzen, die innerhalb dieses Bereichs mehrmals auftauchen. Im distalen Teil der hier untersuchten Sequenz gibt es einen kleineren Bereich mit einem ähnlichen Muster. Die meisten Wiederholungen außerhalb der beiden genannten Bereiche sind sehr kurz und über die gesamte V-Region verstreut. Weiterhin gibt es distal einen besonders langen homologen Bereich, der nur einmal wiederholt wird. Der Pip der Sequenz gegen sich selbst zeigt die Sequenzähnlichkeit der V-Gene untereinander in Form vieler kurzer Wiederholungen mit Identitäten zwischen 50 und 100 % (Anhang I.3). Bemerkenswert sind sich wiederholende Bereiche außerhalb der V-Segmente, die teilweise ebenfalls häufige Wiederholungen zeigen.

Die Sequenz in der Umgebung der D-Segmente ist durch eine besonders ausgeprägte interne Homologie gekennzeichnet (siehe auch Abbildung 2.19). Innerhalb der konstanten Region sind die Ähnlichkeiten im Bereich der C-γ-Gene besonders ausgeprägt, aber auch die Sequenzen der anderen Isotypen zeigen untereinander Ähnlichkeiten.

Abbildung 2.14: Dotplot der Contigs C02 bis C14

Der Dotplot zeigt einen Vergleich der Sequenz mit sich selbst und wurde mit dem Programm Advanced PipMaker erzeugt. Kodierende Sequenzen sind farbig markiert. V-Segmente: rot; D-Segmente: blau; J-Segmente: grün; konstante Regionen: orange (Exon-Bereiche eines Isotyps sind in einer Markierung zusammengefasst). Repetitive Bereiche, die durch den Repeatmasker erkannt wurden, wurden durch den Advanced PipMaker maskiert und sind im Dotplot nicht dargestellt.



2.2.3 Die konstante Region

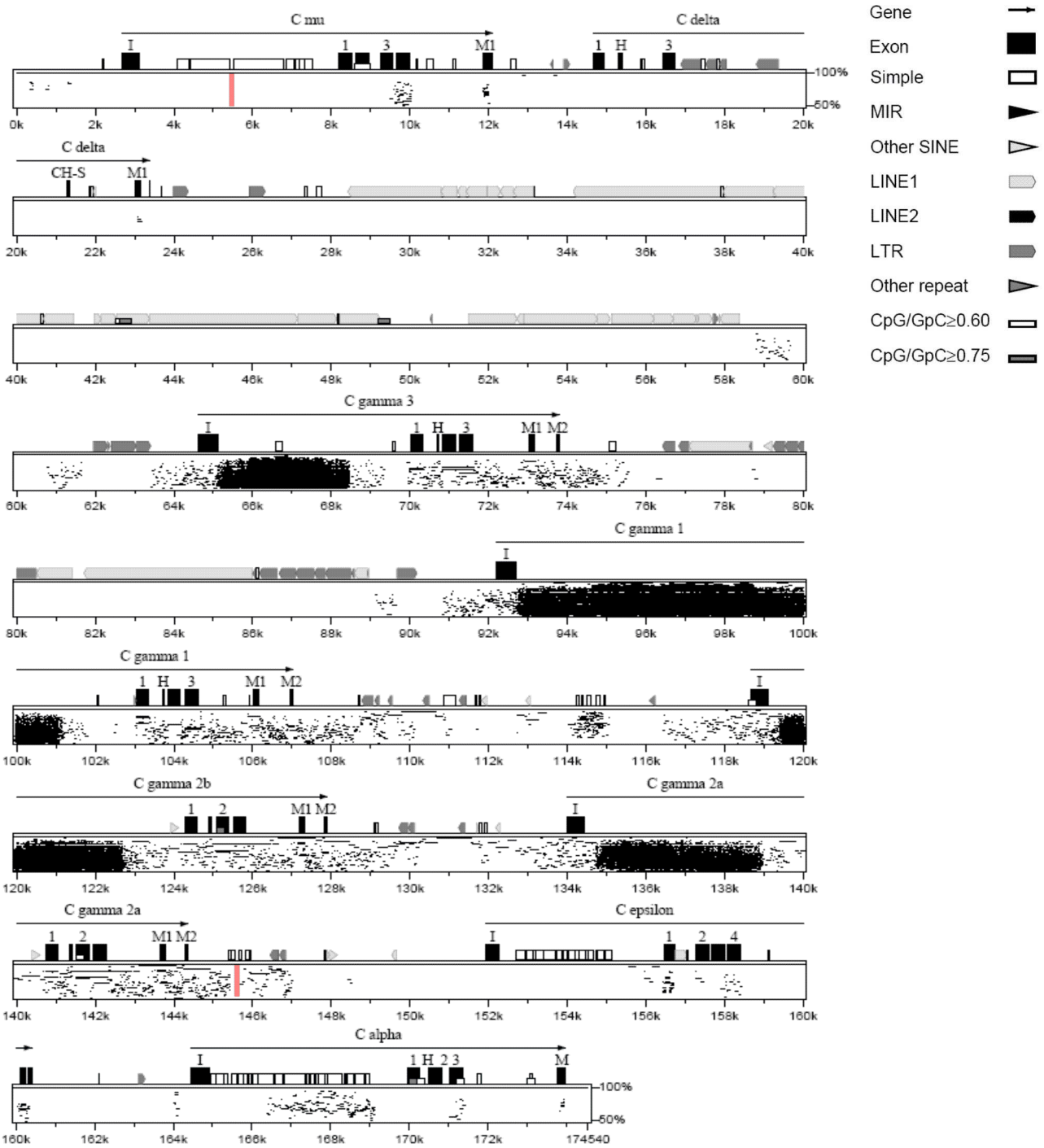
Zur Annotation der konstanten Region wurde die Sequenz der Contigs C12, C13 und C14 mit den bisher veröffentlichten kodierenden Sequenzen und Sequenzannotationen der acht Schwereketten-Isotypen verglichen. Mit Hilfe des MultiPipMakers wurden die Positionen der Exone und Polyadenylierungssignale innerhalb der Contigs ermittelt. Des weiteren wurde die Sequenz mit dem Programm GENSCAN untersucht; GENSCAN ist dafür jedoch nur bedingt anwendbar, da das Programm für die Analyse nicht rekombinierender eukaryotischer Gene entwickelt wurde. Das Ergebnis der Annotation ist in Anhang I.2B dargestellt.

Abbildung 2.15 zeigt die konstante Region als Pip (Percent identity plot) der Sequenz gegen sich selbst mit Markierung der Exone, repetitiven Elemente und Lücken. Die repetitiven Elemente wurden zuvor durch das Repeatmasker-Programm ermittelt. Die S-Regionen werden nur teilweise vom Repeatmasker erkannt: die Sμ-, Sε- und Sα-Regionen werden als 'simple repeats' maskiert. Die Sγ-Regionen werden dagegen vom Repeatmasker nicht erkannt und daher nicht maskiert. Die Sγ-Regionen bestehen vor allem aus 49-bp- und 26-bp-Tandem-Wiederholungen, die die kurzen Wiederholungsmotive der Sμ-Region enthalten [Kataoka et al., 1980; Kataoka et al., 1981; Wu et al., 1984; Mowatt et al., 1985; Szurek et al., 1985; Akahori und Kurosawa, 1997]. Deshalb sind sie im Pip als schwarze Bereiche zu erkennen, da der PipMaker alle Sequenzwiederholungen anzeigt.

Die Sequenz des Contigs C14 endet etwa 200 bp stromabwärts der Polyadenlyierungsstelle des Membran-Exons von Cα. Der 3'-Enhancer ist deshalb nicht mehr enthalten.

Abbildung 2.15: Pip der konstanten Region

Der Pip zeigt den Vergleich der Sequenz der konstanten Region mit sich selbst (AJ851868: Position 1207514 bis Ende). Pfeile fassen die Exone eines Isotyps zusammen, Lücken zwischen den Contigs C12, C13 und C14 sind rot dargestellt; Darstellung der repetitiven Elemente wie in Abbildung I – I-Exon; 1-4 – Immunglobulin-Domänen CH1 bis CH4; H – Hinge-Region; M, M1, M2 – Membran-Exone. Legende der im Pip markierten Elemente:



2.2.4 Die J-Segmente

Im Schwerekettenlocus der Maus gibt es vier J-Segmente, die sich in einem Bereich von 1,38 kb direkt an das DQ52-Element anschließen (Abbildung 2.18 und 2.20). Die J-Segmente von 129/Sv wurden mit den J-Segmenten von BALB/c (EMBL-AC: V00770) und C57BL/6 (NCBI build 34, Chromosom 12 Position 108903384-108906875) verglichen. Die kodierenden Sequenzen der J-Segmente sind bei den drei Mausstämmen mit Ausnahme des JH1-Segments identisch. Das JH1-Segment von 129/Sv kodiert für die gleiche Aminosäuresequenz wie das BALB/c-JH1-Segment (Abbildung 2.16). Beide Segmente unterscheiden sich vom C57BL/6-JH1-Segment durch ein Alanincodon anstatt eines Tyrosincodons in Nukleotidposition 26 bis 28 der kodierenden Sequenz.

Abbildung 2.16: Unterschiede in den JH1-Segmenten von 129/Sv, BALB/c und C57BL/6

Dargestellt ist das Alignment der drei JH1-Segmente mit Übersetzung in Aminosäure-Code in der ersten Zeile ('Translation'). Die Aminosäuresequenz des JH1-Segments von C57BL/6 weicht von den JH1-Segmenten aus 129/Sv und BALB/c in einer Position ab; dies ist in der zweiten Translationszeile dargestellt.

```
Translation      Y   W   Y   F   D   V   W   G   A   G   T   T   V   T   V   S   S
129/Sv       C TAC TGG TAC TTC GAT GTC TGG GGC GCA GGG ACC ACG GTC ACC GTC TCC TCA G
BALB/c       C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --T --- --- -
Translation                                      T
C57BL/6      C --- --- --- --- --- --- --- --- A-- --- --- --- --- --- --- --- --- -
```

2.2.5 Die D-Segmente

Die murinen D-Segmente sind Sequenzen von elf bis 23 bp Länge und kodieren einen Teil der CDR3-Region im Antikörper [Early et al., 1980]. Sie sind umgeben von RSS-Elementen mit 12-bp-Spacer, die für das Rearrangement mit dem V-Segment stromaufwärts und mit dem J-Segment stromabwärts benötigt werden [Sakano et al., 1981; Kurosawa und Tonegawa, 1982]. Die funktionellen D-Segmente werden in vier Familien eingeteilt: Die DQ52-Familie, die DFL-Familie, die DSP-Familie und die DST-Familie [Feeney und Riblet, 1993; Riblet, 2004]. Im Schwerekettenlocus des C57BL/6-Mausstammes wurden von Jian Ye zehn D-Segmente annotiert, die sich jeweils einer der vier Familien zuordnen lassen [Ye, 2004]. Die D-Regionen verschiedener Mausstämme weisen zum Teil erhebliche Unterschiede auf [Trepicchio und Barrett, 1985]. Ye stellt in seiner Untersuchung den Unterschied zwischen C57BL/6 und dem gut untersuchten D-Locus von BALB/c heraus.

Die D-Region unterscheidet sich in 129/Sv deutlich sowohl von C57BL/6 als auch von BALB/c. In der 158-kb-Sequenz zwischen dem nächsten angrenzenden V-Segment musIGHV182 (E4Psi) und dem JH1-Segment wurden fünfzehn D-Segmente identifiziert und jeweils einer der vier D-Segment-Familien zugeordnet. Dabei wurde je ein neues Mitglied der DFL16-Familie und der DST4-Familie beschrieben, DFL16.3 und DST4.3. Die Abbildung 2.17 zeigt die Sequenz der D-Segmente und der RSS-Elemente in 129/Sv als Alignment und nach Familien sortiert. Da das Segment DSP2.2 exakt dupliziert vorliegt, wurden die beiden Segmente zur Unterscheidung der Position mit DSP2.2a und DSP2.2b bezeichnet.

Abbildung 2.17: D-Segmente in 129/Sv

Die Sequenzen der D-Segmente sind nach Familien geordnet und als Aligments dargestellt. Die Position der kodierenden Sequenzen ist gelb markiert, Heptamer- und Nonamer-Sequenzen sind unterstrichen.

```
DFL16.1 GCTTTTTGTGAAGGGATCTACTACTGTGTTTATTACTACGGTAGTAGCTACCACAGTGCTATATCCATCAGCAAAAACC
DFL16.2 ----------------------------C----------.....----------G--A-------A
DFL16.3 -T----G-A---A--C------------A---A----AA--G------CA----A--G-G----T-GT-------GTT

DSP2.2a GATTTTTGTCAAGGGATCTACTACTGTGTCTACTATGATTACGACCACAGTGATATATCCAGCAACAAAAACC
DSP2.2b ----------------------------------------------------------------------
DSP2.3  ----------------------------------------G-------------------------------
DSP2.5  --------------T-------------------------G-A--T--------------------------
DSP2.7  ----------------------------C--------G-A--T--------------------------
DSP2.8  ----------------------------C---G----G-A--T--------------------------
DSP2.9  ----------------------------------TG---G----T--------------------T---------
DSP2.11 ----------------------------C-------AGG------------------------------

DST4    GATTTTGAACAAGTTACTGTCACAGTGAG.ACAGCTCGGGCTACCACTGTAAGAAAAGCTCAAACCAAAACT
DST4.2  ----------------------------G-C---------------------------------------
DST4.3  ------------G--T----------G-G------A-----GT-----A------------------CTG

DQ52    GGTTTTGACTAAGCGGAGCACCACAGTGCTAACTGGGACCACGGTGACACGTGGCTCAACAAAAACC
```

In Abbildung 2.18 ist die D- und J-Region als Pip dargestellt, wobei die 129/Sv-Sequenz mit dem homologen Bereich der C57BL/6-Sequenz verglichen wird. Der Pip zeigt an vielen Stellen eine nahezu hundertprozentige Übereinstimmung der Sequenzen, die jedoch an einigen Stellen unterbrochen wird: so fehlt in C57BL/6 ein LINE1-Element im Bereich 51 bis 57 kb und der Bereich der D-Segmente DST4.3 und DFL16.2. DFL16.3 ist in C57BL/6 vorhanden, wurde jedoch von Ye nicht annotiert.

Abbildung 2.18: Pip der D- und J-Region von 129/Sv im Vergleich mit C57BL/6

Der Pip zeigt die Sequenz von 129/Sv beginnend stromabwärts des JH-proximalen V-Segments musIGHV182 bis zum Beginn des I-Exons von Cμ.. Die C57BL/6-Sequenz wurde aus dem NCBI Mouse build 34 extrahiert (Position 108904000-109049630). Pip-Legende: siehe Abbildung 2.15. Markierung der Segmente: DFL16- blau; DST4 – grün; DSP2 – orange; DQ52 – pink; J- türkis.

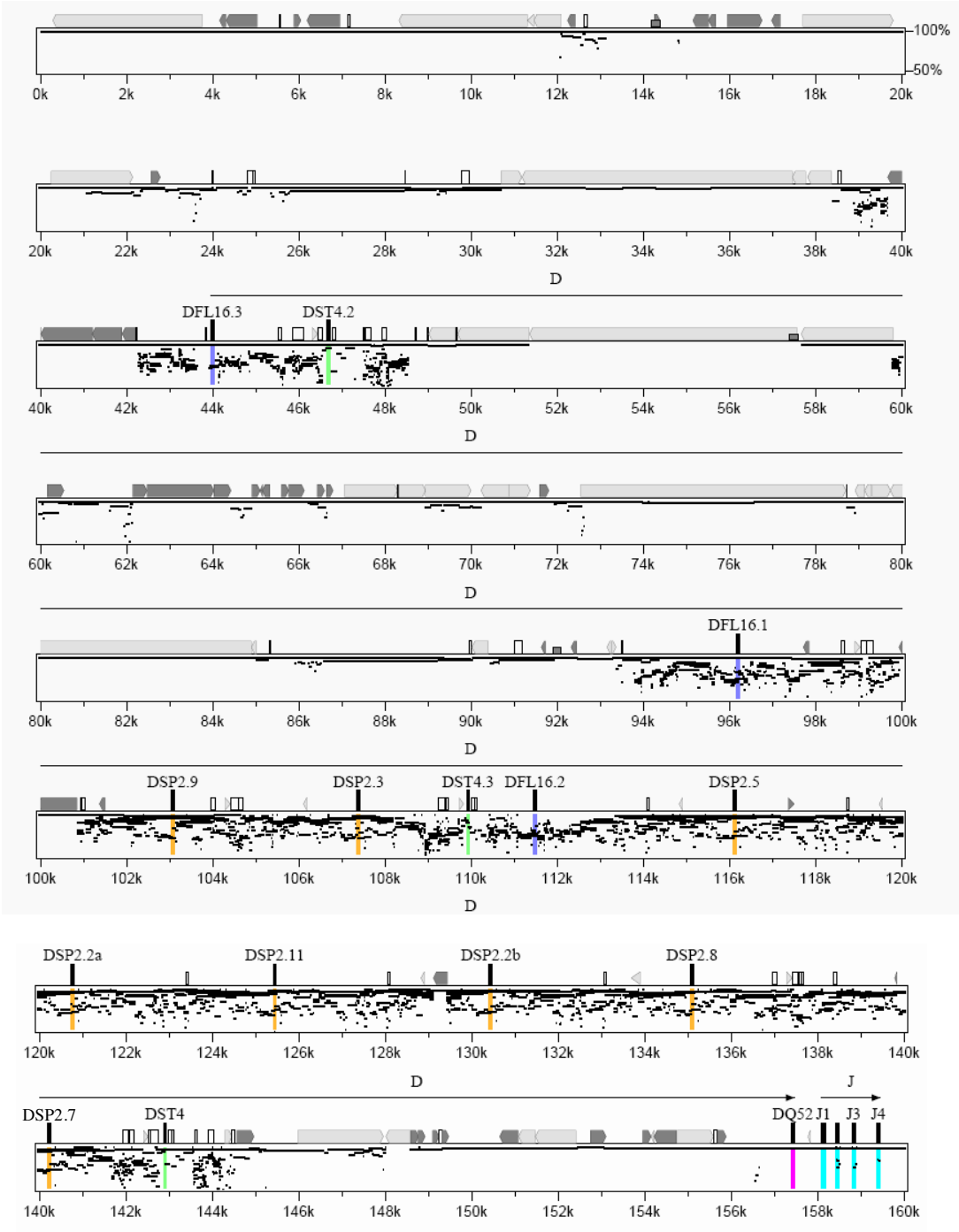
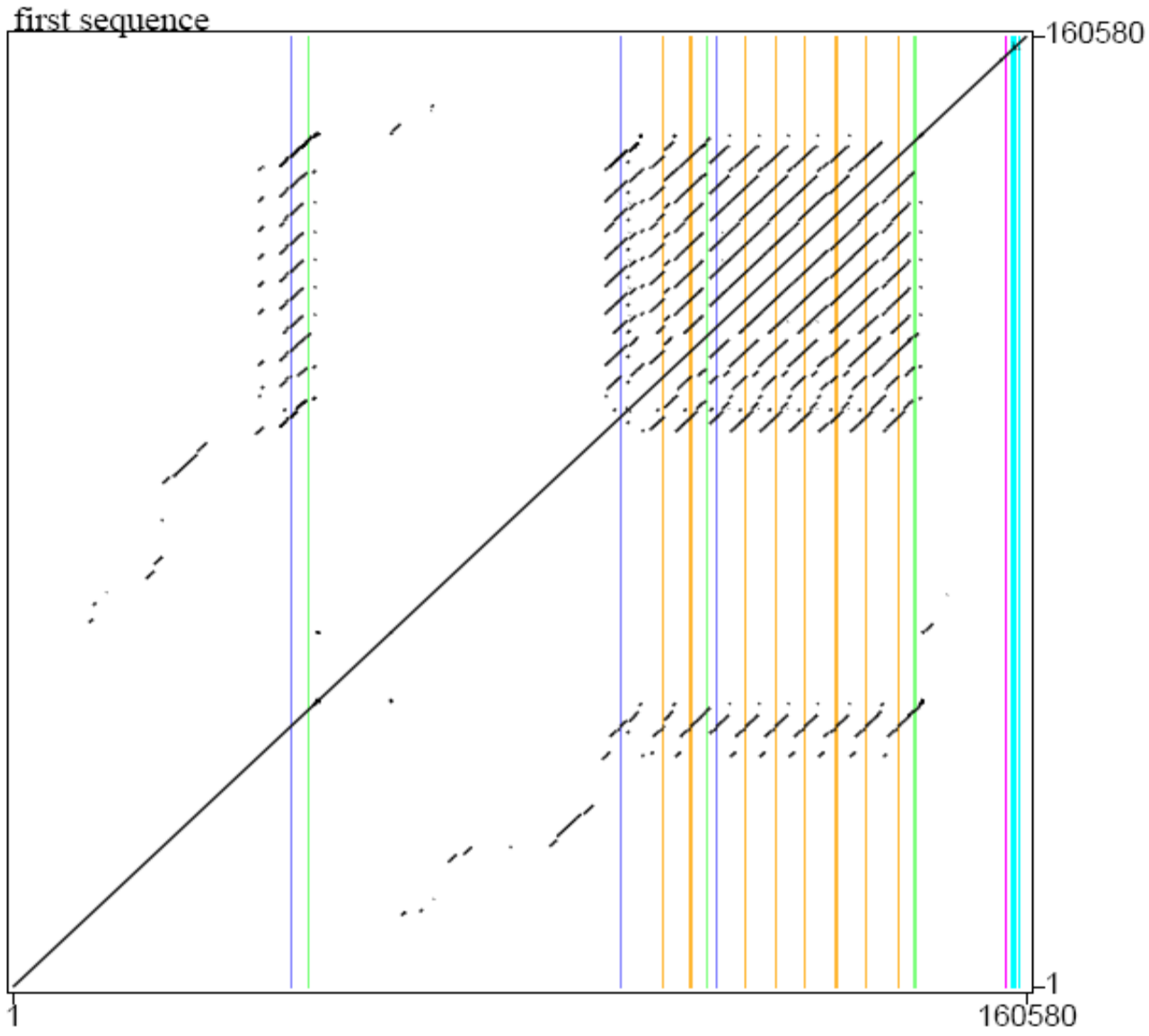


Abbildung 2.19: Dotplot der D- und J-Region von 129/Sv

Der Dotplot zeigt den Vergleich der Sequenz von 129/Sv beginnend stromabwärts des JH-proximalen V-Segments musIGHV182 bis zum Beginn des I-Exons von Cμ mit sich selbst. Markierung der Segmente: DFL16- blau; DST4 – grün; DSP2 – orange; DQ52 – pink; J- türkis.



Sowohl der Dotplot der Contigs C02-C14 (Abbildung 2.14) als auch der Pip der DJ-Region (Abbildung 2.18) zeigen in der D-Region zahlreiche interne homologe Sequenzbereiche. Um diese genauer zu untersuchen, wurde mit dem Advanced Pipmaker ein Dotplot der DJ-Region angefertigt (Abbildung 2.19). Das Ergebnis

zeigt, dass sich die D-Segmente in sich wiederholenden Sequenzabschnitten von mehreren kb befinden. Dies trifft besonders auf den Bereich der DSP2-Familie zu. DQ52 dagegen ist eine unikale Sequenz.

Tabelle 2.13: D-Segmente in den Mausstämmen 129/Sv, C57BL/6 und BALB/c

Angegeben ist das Vorhandensein eines Segments mit '+', das Fehlen mit '-'. Die D-Region von BALB/c ist nicht vollständig sequenziert, so dass keine sicheren Angaben zum Fehlen von Segmenten gemacht werden können (dargestellt als '(-)'). Bei DST4 und DQ52 wurden Haplotypen zugeordnet, weil bei diesen Segmenten sich die Sequenz zwischen den Mausstämmen teilweise unterscheidet, sie aber aufgrund ihrer Position (Ende des DSP-Clusters, Ende der D-Region) eindeutig als Allele identifiziert werden konnten.

* Die Funktionalität dieser Segmente ist nicht nachgewiesen.

D-Segment	129/Sv	BALB/c	C57BL/6
DFL16.1	+	+	+
DSP2.9	+	+	+
DSP2.5	+	+	+
DSP2.2	dupliziert	+	+
DSP2.3, DSP2.4	+	dupliziert	+
DSP2.7	+	+	-
DSP2.8	+	+	-
DFL16.2	+	+	-
DSP2.11	+	+	-
DST4	Igh^a	Igh^a	Igh^b
DQ52	Igh^b	Igh^a	Igh^b
DFL16.3*	+	(-)	+
DST4.2*	+	(-)	+
DST4.3*	+	(-)	-
DSP2.x	-	+	dupliziert
DSP2.6	-	+	-
DSP2.10	-	+	-

Die Tabelle 2.13 zeigt eine Aufstellung der D-Segmente in den Mausstämmen 129/Sv, C57BL/6 und BALB/c. Von insgesamt siebzehn Segmenten sind nur fünf in allen drei Stämmen mit identischer Sequenz vorhanden. Der 129/Sv-Mausstamm teilt insgesamt zehn D-Segmente mit BALB/c und acht D-Segmente mit C57BL/6. Abgesehen von den fünf D-Segmenten, die in allen drei Stämmen vorhanden sind, gibt es nur ein Segment, DSP2.x, das nur in BALB/c und C57BL/6 vorkommt. Da die D-Region des BALB/c-Stammes zwar mit Hilfe von Southern-Blot-Analysen intensiv untersucht worden ist, jedoch keine vollständige genomische Sequenz bekannt ist, kann das Fehlen eines Segments hier nicht sicher festgestellt werden. Ein einfacher Zusammenhang zwischen der Verteilung der Segmente in den drei Mausstämmen läßt sich weder in Hinblick auf die D-Segment-Familie, noch beim Vergleich der genomischen Positionen (hier nicht dargestellt) feststellen. Bemerkenswerterweise gibt es aber kein D-Segment, das nur in C57/BL6 vorkommt.

2.2.6 Die variable Region

2.2.6.1 Annotation funktioneller und nicht-funktioneller V-Segmente

Die variable Region des murinen Schwerekettenlocus wurde bisher nicht vollständig sequenziert. Sie umfasst etwa 3 Mb [Chevillard et al., 2002]. Über die Anzahl der V-Segmente gibt es sehr unterschiedliche Schätzungen [Brodeur und Riblet, 1984; Livant et al., 1986; Gu et al., 1991]. Die aktuelle Schätzung von Riblet beträgt 190 bis 200 vollständige V-Segmente für den Igh^b-Haplotyp; davon wurden 101 V-Segmente durch das Vorhandensein einer mRNA-Sequenz als funktionell identifiziert [Riblet, 2004]. Annähernd übereinstimmend damit veröffentlichte de Bono 2004 eine Analyse der V-Segmente von C57BL/6 mit 104 funktionellen V-Segmenten, 37 Pseudogenen und 74 V-Segment-ähnlichen Sequenzen [de Bono et al., 2004].

Die hier untersuchte Sequenz des 129/Sv-Mausstammes repräsentiert den Igh^a-Haplotyp. Die Contigs C02 bis C11 (Tabelle 2.10) decken den JH-proximalen Teil der variablen Region ab, der sich stromaufwärts an die konstante Region anschließt. Ein weiterer Bereich stromaufwärts von C02 wird von den BACs AJ972403 und AJ972404 abgedeckt. Die V-Segmente dieser BACs wurden annotiert und in Tabelle 2.14 mit aufgeführt, jedoch nicht kartiert, da sowohl die genaue Position der BACs als auch die Position der Fragmente innerhalb der BACs noch unklar ist. Die hier annotierte Sequenz deckt etwas mehr als ein Drittel des gesamten IgH-Locus ab.

Für die Nomenklatur der murinen IgH-V-Gene existiert kein allgemein anerkannter Standard (diskutiert in Kapitel 3.3). In dieser Arbeit werden deshalb behelfsweise die VBASE2-IDs zur Benennung der Segmente eingesetzt.

Mit Hilfe des Prozesses zur automatischen V-Gen-Analyse (siehe Kapitel 2.1.1) wurden in der Sequenz der Contigs C02 bis C11 und der beiden zusätzlichen BACs 117 V-Segmente detektiert und mit VBASE2-IDs benannt. Zur Bewertung der Funktionalität wurden die V-Segmente mit VDJ-Rearrangements aus der EMBL-Bank verglichen. Ebenfalls wurden durch den Prozess V-Gen-Familien und -Namen zugeordnet sowie RSS-Elemente identifiziert. Durch manuelle Annotation wurde der V-Segment-Bereich am 5'-Ende bis zum Beginn des Exons 2 erweitert; die ersten elf Nukleotide des Exons 2 kodieren das Signalpeptid und werden vom V-Gen-Analyse-Prozess nicht berücksichtigt. Das Exon 1, das den größten Teil des Signalpeptids kodiert, wurde in dieser Arbeit nicht annotiert. Eine Tabelle mit dem Annotationsergebnis ist in Anhang I.2A einzusehen. Anhang I.3 zeigt einen Pip der Contigs C02 bis C11, in dem die V-Segmente farbig markiert sind. Abbildung 2.20 zeigt eine maßstabsgetreue Karte des IgH-Locus mit dem Teil der V-Region, welcher von den Contigs C02 bis C11 abgedeckt wird.

Abbildung 2.20: Karte des IgH-Locus im Bereich der Contigs C02 bis C14

Die Abbildung zeigt den bisher sequenzierten Teil der V-Region sowie die D- J- und C-Region des IgH-Locus von 129/Sv. Klasse-1-V-Gene, D- und J-Segmente sowie die Exone der C-Region sind durch lange senkrechte Striche, Klasse-2-V-Gene sind durch kurze senkrechte Striche markiert. Die Dicke eines Strichs deuet die Länge des jeweiligen Elements an. Die Namen der V-Gen-Familien sind farbig markiert, wobei V-Gen-Relikte den Familien zugeordnet wurden, zu dessen Mitgliedern sie die größte Ähnlichkeit aufweisen (siehe auch Tabelle 15). Die Namen der V-Gen-Relikte sind grau hinterlegt.

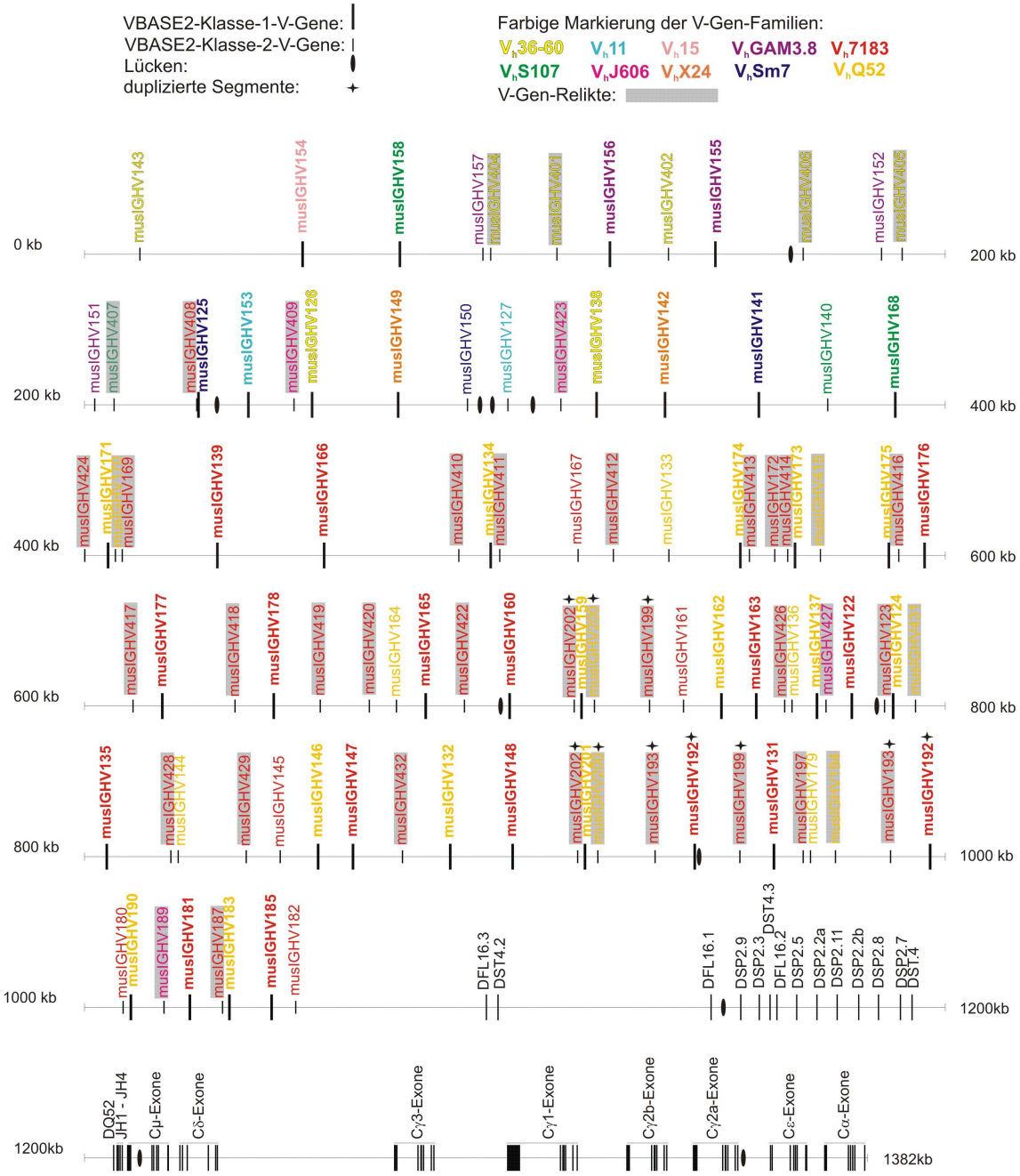


Tabelle 2.14: V-Gene im proximalen Teil des IgH-Locus von 129/Sv

V-Gen-Familie	Summe der V-Segmente	Rearrangement bekannt	Funktionalität nicht bekannt	Pseudogen
V_h7183	22	17	2	3
V_hQ52	19	14	3	2
V_hS107	3	2	-	1
V_hSm7	3	2	-	1
V_hX24	2	2	-	0
V_h11	2	1	1	0
V_h36-60	6	4	-	2
V_hGAM3.8	6	2	4	0
V_h3609N	1	0	1	1
V_h12	1	1	-	0
V_hJ606	5	2	3	0
V_h15	1	1	-	0
nicht klassifiziert	1	0	1	0
Relikte	45	0	-	45
Summe	117	48	15	55

Die Tabelle 2.14 zeigt eine Übersicht über die Anzahl der hier annotierten V-Gene und -Familien, inklusive der BACs AJ972403 und AJ972404. Die V-Gene des murinen IgH-Locus liegen in Clustern von V-Genen gleicher Familien vor, wobei benachbarte Cluster teilweise durchmischt sind [Review: Kofler et al., 1992]. Den größten Anteil an V-Genen in der hier untersuchten Sequenz nehmen die Familien V_h7183 und V_hQ52 mit 22 beziehungsweise neunzehn Mitgliedern ein. Alle anderen Familien kommen im vorliegenden Abschnitt mit ein bis sechs Segmenten vor. Die Familien V_h10, V_hJ558 und V_h3609P fehlen vollständig, da sie sich im JH-distalen Teil des Locus befinden [Mainville et al., 1996].

48 V-Segmente wurden in Rearrangements der EMBL-Bank identifiziert und der VBASE2-Klasse 1 zugeordnet. Die Existenz eines funktionellen DNA-

Rearrangements beweist zwar noch nicht das Vorkommen eines V-Segments in einem funktionellen Antikörper, ist jedoch ein starker Hinweis auf die Funktionalität. Für 69 V-Segmente wurde kein passendes Rearrangement gefunden, so dass sie der VBASE2-Klasse 2 zugeordnet wurden. Sechzehn davon zeigen keine offensichtlichen Defekte in der kodierenden Sequenz und lassen sich eindeutig einer V-Gen-Familie zuordnen. Zwei dieser Segmente, musIGHV167 und musIGHV136, weichen im RSS-Element von der Konsensus-Sequenz ab und wurden deshalb als Pseudogene eingestuft. Die übrigen vierzehn V-Gene dieser Gruppe sind möglicherweise funktionell; auch durch manuelle BLAST-Suchen der Segmente in der EMBL-Bank konnte jedoch zu keinem eindeutig ein Rearrangement zugeordnet werden, so dass sie, wenn man Funktionalität annehmen möchte, zumindest als selten gebrauchte V-Segmente einzustufen sind.

Ein Klasse-2-V-Segment, musIGHV402, läßt sich keiner Familie eindeutig zuordnen, zeigt jedoch auch keine offensichtlichen Defekte. Das RSS-Element unterscheidet sich nur in der letzten Position des Nonamers von RSS-Elementen der V_h36-60-Familie. Eine BLAST-Suche des Segments in der EMBL-Bank zeigt eine sehr starke Ähnlichkeit zu einigen patentierten humanisierten Antikörpern (zum Beispiel EMBL-ACs A43069, I64624). In C57BL/6 kommen ähnliche Sequenzen in den BACs AC073589 und AC073563 vor. Da es aber keinen eindeutigen Hinweis auf die Funktionalität von musIGHV402 in der Maus gibt, wird es als 'nicht klassifiziert' bezeichnet.

2.2.6.2 V-Gen-Relikte

54 Klasse-2-V-Segmente tragen Stop-Codons oder haben andere Defekte in der kodierenden Sequenz, die eine Funktionalität der Segmente verhindern. Davon

haben 45 Segmente weniger als 70% Identität mit einem Klasse-1-V-Gen und lassen sich insofern keiner Familie zuordnen, da murine IghV-Gene einer Familie per Definition mindestens zu achtzig Prozent identisch sind [Review: Kofler et al., 1992]. Diese Segmente werden in der vorliegenden Arbeit als V-Gen-Relikte bezeichnet.

Tabelle 2.15: V-Gen-Relikte und zugeordnete Klasse-1-V-Segmente

Klasse-1-Segment	Familie	Relikte
musIGHV114	V_hJ606	musIGHV409 musIGHV189, musIGHV427, musIGHV400, musIGHV423
musIGHV128	V_h36-60	musIGHV405, musIGHV401, musIGHV402, musIGHV404
musIGHV138	V_h36-60	musIGHV406
musIGHV139	V_h7183	musIGHV169, musIGHV419
musIGHV148	V_h7183	musIGHV432, musIGHV420, musIGHV429, musIGHV426, musIGHV428, musIGHV410, musIGHV411, musIGHV417, musIGHV414, musIGHV199, musIGHV202, musIGHV172, musIGHV193, musIGHV187
musIGHV154	V_h15	musIGHV399
musIGHV158	V_hS107	musIGHV407
musIGHV160	V_h7183	musIGHV123, musIGHV197, musIGHV408, musIGHV412
musIGHV166	V_h7183	musIGHV416
musIGHV175	V_hQ52	musIGHV431, musIGHV415, musIGHV194, musIGHV200, musIGHV170
musIGHV178	V_h7183	musIGHV413, musIGHV424
musIGHV181	V_h7183	musIGHV418, musIGHV422

Ein Vergleich der Relikte mit den V-Segmenten der Klasse 1 zeigt, dass es innerhalb der Relikte Gruppen ähnlicher Sequenzen gibt, die gemeinsam die gleiche Sequenz als besten Treffer in Klasse 1 aufweisen. So lassen sich die 45 Relikte zwölf Klasse-1-V-Genen zuordnen. Auch musIGHV402 läßt sich hier in eine Gruppe der 36-60-Familie einordnen. Die Relikt-Gruppen und die

entsprechenden Klasse-1-Segmente sind in Tabelle 2.15 aufgeführt. 73% der Relikte (33 Segmente) haben demnach Ähnlichkeit mit der V_h7183-Familie.

2.2.6.3 RSS-Elemente

Bei den RSS-Elementen am 3'-Ende der V-Gene ist eine Heptamer- und eine Nonamer-Konsensus-Sequenz durch einen 23 +/- 1 bp langen Spacer getrennt [Review: Tonegawa, 1983]. Obwohl sich für Heptamer und Nonamer eine allgemeine Konsensus-Sequenz bestimmen lässt, weisen die Sequenzen eine erhebliche Heterogenität auf. Nur die ersten drei Nukleotide des Heptamers sowie die 5. und 6. Position des Nonamers sind besonders stark konserviert [Ramsden et al., 1994].

Bei der Untersuchung der 3'-Enden der 117 V-Segmente wurde bei 72 Segmenten ein vollständiges oder verkürztes RSS-Element identifiziert. Die 45 Segmente ohne erkennbares RSS-Element sind Pseudogene oder V-Gen-Relikte. Es gibt auch einige V-Gen-Relikte mit einem RSS-Element. Die Heptamer- und Nonamer-Sequenzen wie auch die dazwischen befindlichen Spacer zeigen starke Ähnlichkeiten innerhalb der V-Gen-Familien [Schroeder et al., 1990; Review: Feeney et al., 2000]. Dies konnte für die hier untersuchten RSS-Elemente bestätigt werden (Daten nicht dargestellt). De Bono et al. stellte 2004 einen 22 bp Spacer für alle V-Segmente der V_hJ558-, V_hQ52- und V_hS107-Familie in C57BL/6 fest [de Bono et al., 2004]. Demgegenüber weisen die RSS-Elemente der V_hQ52- und V_hS107-Familie in 129/Sv 23 bp Spacer auf (Abbildung 2.21). Sieben der V_hQ52-Elemente haben in der sechsten Position des Nonamers abweichend von der Konsensus-Sequenz ein Thymin-Nukleotid, sind aber dennoch der Klasse 1 zugeordnet; die Abweichung von der

Konsensus-Sequenz verhindert in diesem Fall also nicht die Bildung eines VDJ-Rearrangements.

Abbildung 2.21: Alignment der RSS-Elemente der Familien V_hQ52 und V_hS107 in 129/Sv

Die Familie und VBASE2-Klasse sind, durch Punkte getrennt, der VBASE2-ID angefügt; Pseudogene sind mit 'psi' markiert. Nicht dargestellt sind musIGHV136 (V_hQ52) musIGHV140 (V_hS107), die kein Konsensus-Nonamer aufweisen.

RSS Konsensus	CACAGTGXXXXXXXXXXXXXXXXXXXXXACAAAAACC
musIGHV171.Vh-Q52.1	-------AGGGAAGTCCAGTGTGAACTTGC---------
musIGHV124.Vh-Q52.1	-------AGGGAAGTCCATTATGAACCTAA-------TT
musIGHV144.Vh-Q52.2.psi	-------AGAGAAGTCCATTATGAACCTAA-------TT
musIGHV159.Vh-Q52.1	-------AGGGAAGTCCATTATGAACCTGA-------TT
musIGHV201.Vh-Q52.1	-------AGGGAAGTCCATTATGAACCTGA-------TT
musIGHV173.Vh-Q52.1	-------AGGGAAGTCCATTATGAACTTGA-------TT
musIGHV183.Vh-Q52.1	-------AGGGAAGTCCATTATGAACTTGA-------TT
musIGHV134.Vh-Q52.1	-------AGGGAAGTCCAATGTGAGCCTGC-----T---
musIGHV175.Vh-Q52.1	-------AGGGAAGTCCAATGTGAGCCTGC-----T---
musIGHV132.Vh-Q52.1	-------TGGGAAGTCCAATGTGAGCCTGC-----T--T
musIGHV146.Vh-Q52.1	-------TGGGAAGTCCAATGTGAGCCTGC-----T--T
musIGHV162.Vh-Q52.1	-------TGGGAAGTCCAATGTGAGCCTGC-----T--T
musIGHV137.Vh-Q52.1	-------AGGGAAGTCCAGTGTGAGCCTGC-----T--T
musIGHV164.Vh-Q52.2	-------AGGGAAGCCCAGTGTGAGCCTGC-----T--T
musIGHV133.Vh-Q52.2	-------AGAGAAGTCCAGTGTGAGCATTC-----T--T
musIGHV174.Vh-Q52.1	-------TGGGAAGTCCAATGTGAGCATTC-----T--T
musIGHV179.Vh-Q52.2.psi	----C--AGGGAAGTCCATTATGAACTTGA-------TT
musIGHV190.Vh-Q52.1	----A--AGAGAAGTCCATTGTGAGCATTC-----T--T
musIGHV168.Vh-S107.1	-------AGAGGACGTCATTGTGAGCCCAG---C-----
musIGHV158.Vh-S107.1	-------AGGGTACTTCAGTGTGAGCCTAG---C-----

2.2.6.4 Exakte Duplikationen von V-Segmenten

Fünf V-Segmente im Bereich der V_h7183- und V_hQ52-Familie kommen als exakte Duplikationen ein zweites Mal vor: musIGHV192, musIGHV193, musIGHV199, musIGHV200 und musIGHV202 (Tabelle 2.16). Vier dieser duplizierten Segmente sind V-Gen-Relikte, eines wird der VBASE2-Klasse 1 zugeordnet. Eines der Relikte stammt von der V_hQ52-Familie ab, die anderen Duplikationen gehören zur V_h7183-Familie. Alle fünf Segmente sind auch in C57BL/6 enthalten (Tabelle 2.16), im NCBI-Assembly sind sie jedoch nicht dupliziert [de Bono, 2004; Riblet, 2004]. Die C57BL/6-Segmente Vh7183.b6 und Vh7183.b7psi sowie Vh7183.b10psi, VhQ52b5.psi und Vh7183.b11psi (Nomenklatur nach Riblet) liegen jeweils in direkter Nachbarschaft.

Tabelle 2.16: V-Gen-Duplikationen im V_h7183/V_hQ52-Bereich

[1] Riblet, 2004; [2] de Bono et al., 2004; 3 Williams et al., 2001; [4] IMGT germline gene and allele table (http://imgt.cines.fr:8104).

VBASE2-ID	erste Position	zweite Position	Familie	Namen
musIGHV192	941778-942084	996469-996775	V_h7183	Vh7183.b6[1], 5-3b-5[2], 3:3.9[3], IGHV5S6*01[4]
musIGHV193	932597-932901	987288-987592	V_h7183	Vh7183.b7psi[1]
musIGHV199	952322-952626	731327-731631	V_h7183	Vh7183.b10psi[1]
musIGHV200	919344-919644	718542-718842	V_hQ52	VhQ52.b5psi[1]
musIGHV202	713774-714075	914562-914863	V_h7183	Vh7183.b11psi[1]

3. Diskussion

3.1 Automatische Annotation von Immunglobulinen

Die bisherigen Sammlungen von Keimbahn-V-Gen-Sequenzen basieren auf manuellen Datenbank- und Literatur-Recherchen. Die manuelle Annotation und Datenbank-Pflege gewährleistet eine hohe Qualität der Datenbank-Einträge und ist ohne Zweifel der automatischen Annotation in vielerlei Hinsicht überlegen. Die Automatisierung von Prozessen ist jedoch wichtig und sinnvoll, wenn die Bearbeitung großer Datenmengen erforderlich ist und wenn ein Teil der manuellen Arbeit ohne Qualitätsverlust durch Programme ersetzt werden kann. Mit dem Ziel einer effizienten und qualitativ hochwertigen Datenbearbeitung wurde in dieser Arbeit so viel wie möglich automatisiert und so wenig wie nötig manuell annotiert.

Die Automatisierung der V-Gen-Analyse ist Grundlage des Konzeptes der VBASE2-Datenbank. Die Bewertung der EMBL-Bank-Einträge im Analyse-Prozess erfolgt aufgrund der Sequenz, so dass der Prozess unabhängig von der Qualität der Annotation des EMBL-Eintrags abläuft. Dabei sind die Randbedingungen und Vorlagen-Sequenzen, die für die automatische Analyse nötig sind, durch visuelle Sequenzauswertung entstanden. So ist eine grundlegende Voraussetzung für den Prozess die Kenntnis über zumindest einige Keimbahn-V-Gene, deren Sequenz das gesamte Repertoire in hinreichender Weise repräsentiert. Auch die DNAPLOT-Master-Sequenzen, die als Familien-Konsensus-Sequenzen fungieren und die Lückenmuster für eine einheitliche Numerierung der variablen Region enthalten, sind erforderlich, ebenso die Sequenzen der J-Segmente und RSS-Elemente. Weiterhin geht in den Filter für synthetische Sequenzen eine manuell gepflegte Liste bekannter synthetischer Antikörper ein. Der Prozess als solcher muss also manuell gepflegt werden. Auf der anderen Seite ist zu berücksichtigen, dass auch bei

manueller Annotation eines unbekannten V-Gens die eigentliche Sequenzanalyse gewöhnlich einem Programm überlassen wird, das die zu untersuchende Sequenz mit bekannten Sequenzen vergleicht. Die Aufgabe des automatischen Prozesses besteht also vor allem darin, die erforderlichen Analyseschritte nacheinander durchzuführen und entsprechend der Ergebnisse eine Klassifizierung vorzunehmen.

Die sequenzbasierte V-Gen-Analyse ist für die Automatisierung sehr gut geeignet, weil sich exakte Bedingungen für die V-Gen-Bewertung formulieren lassen. So können zum Beispiel V-Gen-Rearrangements eindeutig an der Existenz eines J-Segments stromabwärts des jeweiligen V-Gens identifiziert und Keimbahn-Konfigurationen am Vorhandensein eines RSS-Element erkannt werden. Auf diese Art wird die Einordnung eines V-Gens in die VBASE2-Klasse 1 durch mehrere unabhängige Tests abgesichert. Eine wichtige Voraussetzung für die eindeutige automatische Zuordnung ist die Tatsache, dass biologische Sequenzen nicht zufällig verteilt sind. Zufällige Sequenzen würden möglicherweise zum Abdriften des Datensatzes führen, da das Prozessergebnis die Anfangsbedingungen für den nächsten Prozessdurchlauf definiert und die BLAST-Suche mit jedem Durchlauf mehr potentielle V-Gene entdecken würde, die den Ursprungssequenzen immer unähnlicher wären. Stattdessen zeigt das Ergebnis der BLAST-Suche von Klasse-2-Sequenzen (Tabelle 2.4), dass die Gruppe der V-Gen-Sequenzen eines Locus klar begrenzt ist und es keine Sequenzen gibt, die dieser Gruppe von V-Genen nur „ein bisschen“ ähnlich sind. Die vorhandenen Sequenzen lassen sich eindeutig zuordnen in „V-Gen dieser Gruppe“ und „Nicht V-Gen dieser Gruppe“.

Wesentlicher Vorteil der Automatisierung ist die einfache regelmäßige Aktualisierung des Datensatzes. Diese ist insbesondere erforderlich, um die

Verbindung von VBASE2 mit Ensembl zu ermöglichen. Ensembl bietet über den DAS Server den Vorteil, ein etabliertes System zur Präsentation der V-Gene im Kontext der genomischen Umgebung zu nutzen. Da jedoch die Sequenzierung der Genome von Maus und Mensch noch nicht abgeschlossen ist, ändern sich die Sequenzen und damit die Positionen der V-Gene in Ensembl regelmäßig. Eine Verbindung zu Ensembl ist dementsprechend nur möglich, wenn der eigene Datensatz stets an die neuen Ensembl-Sequenzen angepasst wird. Ein weiterer wichtiger Aspekt der Automatisierung ist die Tatsache, dass die Automatisierung die Pflege der Datenbank mit minimalem personellen Aufwand erlaubt und damit den Bestand und die Aktualität der Datenbank auch für die Zukunft sichert. Die Automatisierung erlaubt außerdem, mit jedem Prozessdurchlauf die Datenbank vollständig neu zu erzeugen. Dies ermöglicht uneingeschränkte Flexibilität in Bezug auf das Datenbank-Schema.

Die Automatisierung setzt der V-Gen-Analyse allerdings auch Grenzen. Wie bereits bei der Beschreibung der Parameter-Optimierung in Kapitel 2.1.1.3.1 angesprochen, ist eine hohe Stringenz zur sicheren Detektion eines Keimbahn-V-Gens erforderlich, die die Zahl der verwendbaren EMBL-Bank-Einträge und damit die Zahl der resultierenden Keimbahn-V-Gene limitiert. Des weiteren wird die CDR3-Region bisher von der automatischen Analyse ausgeschlossen, da hier durch den Rekombinationsprozess Sequenzveränderungen entstehen. Es ist theoretisch denkbar, dass zwei verschiedene Keimbahn-V-Gene existieren, die sich nur im CDR3-Bereich unterscheiden. Diese würden vom Analyse-Prozess als identische V-Gene detektiert werden.

Neben der Erzeugung einer V-Gen-Datenbank gibt es weitere Anwendungsmöglichkeiten für die automatische V-Gen-Analyse. Der wichtigste Anlass für die Entwicklung des Analyse-Prozesses war die in Kapitel 2.2.6

beschriebene Annotation der variablen Region des IgH-Locus der 129/Sv-Maus, die durch den Prozess ganz erheblich unterstützt wurde. Auch in anderen Datenbanken und Informationssystemen konnte und kann der Prozess in Zukunft Anwendung finden. In Kapitel 2.1.3 wird beschrieben, wie ein Teil des Analyse-Prozesses so modifiziert wurde, dass die Immunglobuline in der UniProtKB/TrEMBL-Datenbank damit kontrolliert werden können. Weiterhin ist in Kooperation mit der Ensembl-Arbeitsgruppe am Sanger Centre ein Projekt geplant, in dem der Analyse-Prozess in die Form eines Moduls der Ensembl Annotationspipeline gebracht werden soll. Ziel ist es, die automatische Annotation der Immunglobuline in Ensembl erheblich zu verbessern, die wegen der ungewöhnlichen Genstruktur durch die Standard-Methoden der Annotationspipeline nur unzureichend erfolgt. Auch eine Kooperation mit der Arbeitsgruppe des Vega Browsers am Sanger Centre ist in Planung. Diese Gruppe arbeitet an der manuellen Genom-Annotation, die wegen der enormen Mengen an zu bearbeitenden Daten nur sehr langsam voranschreitet. Für die Annotation unbekannter Immunglobulinloci wäre eine automatisch erzeugte Datengrundlage als Vorbereitung der manuellen Annotation sehr vorteilhaft.

Der in dieser Arbeit entwickelte Analyse-Prozess ist nicht der einzige Ansatz zur Automatisierung der V-Gen-Analyse. De Bono und Chothia stellten 2003 die Exegesis-Prozedur vor, die die automatische Annotation von Proteinen der Immunglobulin-Superfamilie (IgsF) in Ensembl deutlich verbessern kann [de Bono und Chothia, 2003]. Exegesis wurde als Grundlage für eine vergleichende Analyse von IgH-V-Genen in Maus und Mensch genutzt, die de Bono 2004 veröffentlichte [de Bono et al., 2004]. Um V-Gene zu detektieren, die von der Ensembl Pipeline nicht annotiert und daher durch Exegesis nicht erfasst werden, wurden hidden Markov models (HMM) von IgSF-Sequenzen auf die Sequenz des murinen gH-Locus in Ensembl angewendet. Das Ergebnis sind 141

V-Gen-Sequenzen, von denen 104 durch manuelle Untersuchung als funktionell eingestuft werden. Alle 141 Sequenzen sind auch in VBASE2 enthalten, allerdings gehören weniger als 104 der Klasse 1 an. De Bono hat bei seiner manuellen Prüfung der Funktionalität auch Mutationen in den VDJ-Rearrangements zugelassen, die im automatischen Prozess nicht zugeordnet werden können. VBASE2 enthält insgesamt 178 V-Gene mit einer Referenz zu Ensembl, davon gehören 63 der Klasse 1 und 115 der Klasse 2 an. Die von de Bono angewendeten HMMs zur V-Gen-Detektion erfordern einen höheren Rechenaufwand als die im VBASE2-Prozess verwendetete BLAST-Suche, und der Prozess von de Bono benötigt insgesamt mehr Zeit und personellen Einsatz. Er ist daher für die Integration in vollständig automatisierte Prozesse wie die Generation von UniProtKB/TrEMBL nicht geeignet.

3.2 Die V-Gen-Datenbank VBASE2

3.2.1 Die Strategie von VBASE2

Da bereits mehrere öffentliche V-Gen-Datenbanken zur Verfügung stehen, ist die Frage offenkundig, warum in dieser Arbeit eine weitere V-Gen-Datenbank entwickelt wurde. Für den Bedarf an einer neuen V-Gen-Datenbank gibt es mehrere Gründe: Wie bereits in der Einleitung und in Kapitel 2.1.1.3.2 besprochen, sind die Keimbahn-Datensätze der etablierten Datenbanken nicht identisch. Weiterhin besteht teilweise Unklarheit darüber, nach welchen Kriterien die Keimbahn-V-Gene als solche identifiziert wurden, wodurch ihre Funktionalität belegt ist und wann die manuell erstellten Datensätze aktualisiert werden. Die etablierten V-Gen-Datenbanken sind untereinander nicht vernetzt, und es besteht keine Verbindung zur genomischen Sequenz der Loci in Ensembl.

Mit der Entwicklung von VBASE2 ist erstmals eine systematische Klassifizierung aller V-Gen-Sequenzen der EMBL-Bank vorgenommen worden. VBASE2 verweist auf Ensembl, IMGT/LIGM-, Kabat- und Vbase-Einträge und zeichnet sich durch einen transparenten Informationsfluss aus. Der VBASE2-Datensatz wird im Abstand von ein bis drei Monaten vollständig aktualisiert.

In den VBASE2-V-Gen-Einträgen werden alle Quellen für eine V-Gen-Sequenz in der EMBL-Bank und in Ensembl angegeben. Die Art und Anzahl der Originalsequenzen liefert wichtige Informationen über Zuverlässigkeit, Bekanntheitsgrad und Literaturreferenzen zu einem V-Gen. Weiterhin verweist VBASE2 auf Einträge der IMGT/LIGM-Datenbank, um die zum Teil detaillierte manuelle Annotation einiger IMGT/LIGM-Einträge dem Nutzer zur Verfügung zu

stellen. Ebenso gibt VBASE2 die Kabat-IDs und Vbase-IDs der V-Gene an, da viele Nutzer mit diesen IDs zur Benennung von V-Genen vertraut sind. Auch in der V-Gen-Nomenklatur schafft VBASE2 mehr Transparenz durch Angabe aller bekannten Trivialnamen eines V-Gens. So werden zum Beispiel die unterschiedlichen Nomenklaturen von de Bono und von Riblet, die 2004 unabhängig voneinander eine Karte der Igh-V-Gene in C57BL/6 veröffentlichten, nebeneinander aufgeführt.

3.2.2 Der V-Gen-Datensatz von VBASE2

Die Qualität einer Keimbahn-V-Gen-Datenbank wird vor allem durch zwei Kriterien bestimmt: Zum einen müssen alle V-Gene der Datenbank in der exakten Sequenz tatsächlich im vererbten Genom vorkommen, zum anderen sollen möglichst alle bekannten Keimbahn-V-Gene im Datensatz enthalten sein. In VBASE2 wird durch die Einteilung der Sequenzen in drei Klassen innerhalb der Datenbank eine Qualitätsunterscheidung vorgenommen. Da Klasse-1-Sequenzen durch mindestens ein Rearrangement und eine identische Keimbahn-Sequenz belegt sind, ist die Wahrscheinlichkeit des Auftretens von Sequenzierfehlern und somatischen Mutationen in dieser Klasse relativ gering. Sequenzen der Klasse 2 sind dagegen teilweise nur durch eine einzige Sequenz aus einem nicht abgeschlossenen Sequenzierprojekt belegt, so dass hier durchaus Sequenzierfehler auftreten können. In Klasse 3 besteht die Möglichkeit, dass somatische Mutationen vorhanden sind, die in unterschiedlichen Rearrangements in denselben Positionen auftreten. So ist bekannt, dass bei der Affinitätsreifung von Antikörpern gegen das Hapten phOx (2-Phenyl-Oxazolon) bestimmte Schlüsselmutationen gehäuft auftreten, die für

die Bildung hochaffiner Antikörper verantwortlich sind [Berek et al. 1991]. Um identische Mutationen in verschiedenen Rearrangements eines Klasse 3-Eintrags zu vermeiden, wird derzeit ein textbasierter Filter verwendet, der mehrfache Klasse-3-Referenzen aus derselben Arbeitsgruppe ausschließen soll. Als Merkmal zur Unterscheidung der veröffentlichenden Arbeitsgruppe wird dabei die Autorenzeile des EMBL-Bank-Eintrags verwendet. Diese Text-Mining-Funktion muss jedoch noch verbessert werden, weil eine Arbeitsgruppe häufig verschiedene EMBL-Bank-Veröffentlichungen hat, bei denen sich der Wortlaut der Autorenzeile unterscheidet. Weiterhin könnten in Klasse 3 dsFv-Fragmente als Referenzen auftreten, die durch die manuell gepflegte Liste synthetischer Antikörper bisher nicht erfasst wurden und deshalb im Analyse-Prozess nicht herausgefiltert werden. Auf der anderen Seite sind in den Klassen 2 und 3 viele V-Gene enthalten, die tatsächlich in der Keimbahn vorkommen und teilweise auch funktionell sind. So zeigt der Vergleich der murinen IgH-V-Gene von VBASE2 mit der Annotation von de Bono, dass 41 von de Bono als funktionell eingestufte V-Gene der VBASE2-Klasse 2 angehören. Die Klassen 2 und 3 erlauben es also, dem Nutzer auch solche Sequenzen zur Verfügung zu stellen, deren Zuverlässigkeit im automatischen Prozess nicht sicher festgestellt werden kann. Damit tragen diese Klassen wesentlich zur Vollständigkeit des Datensatzes bei, ohne gleichzeitig die Qualität der gesamten Datenbank zu gefährden. Durch den Verweis auf die Originalsequenzen kann der Nutzer die Qualität der jeweiligen V-Gen-Sequenz stets selbst überprüfen.

Die in Tabelle 2.5 aufgestellte Übersicht über die Anzahl der Keimbahn-V-Gene von VBASE2, Vbase, Almagro und vom IMGT zeigt, dass auch VBASE2 zur Zeit noch keine vollständige Keimbahn-V-Gen-Sammlung präsentieren kann. Der große Unterschied zwischen Vbase und VBASE2 vor allem am humanen IgH-Locus deutet an, an welchen Stellen die automatische Annotation gegenüber

der manuellen Analyse Lücken aufweist. So sind in Vbase auch V-Gen-Allele von IgM-Antikörpern enthalten, die gewöhnlich nicht mutiert sind und sich deshalb durch ein einziges Rearrangement belegen lassen. Diese Sequenzen können im automatischen Prozess nicht berücksichtigt werden. Gerade bei den humanen Sequenzen, die auch die Ergebnisse diagnostischer Untersuchungen von Patienten enthalten und deshalb viele verschiedene Haplotypen repräsentieren, ist die manuelle Analyse daher offensichtlich ergiebiger. Allerdings wurde 1998, also nach der Erstellung von Vbase, bekannt, dass es IgM-Antikörper von humanen Gedächtnis-B-Zellen gibt, die somatische Mutationen tragen [Klein et al., 1998]. Unter diesem Gesichtspunkt ist die doppelte Absicherung von Sequenzen eben auch ein Qualitätsmerkmal der VBASE2-Klasse-1-Sequenzen, und es ist im Einzelfall zu prüfen, ob es sich bei den V-Genen, die in VBASE2 fehlen, um zuverlässige Keimbahn-Sequenzen handelt.

Bei der Betrachtung der V-Gen-Sequenzen, die nur in VBASE2 enthalten sind, in den anderen V-Gen-Sammlungen dagegen fehlen, gibt es große Unterschiede zwischen den Datensätzen von Maus und Mensch sowie zwischen den VBASE2-Klassen, denen diese Sequenzen angehören (Tabelle 2.6). Beispielsweise gibt es nur eine einzige IgH-V-Klasse-1-Sequenz, die in Vbase nicht enthalten ist. Dies ist dadurch erklärbar, dass der humane IgH-Locus zur Zeit der Erstellung von Vbase bereits vollständig sequenziert worden war, so dass in den letzten Jahren kaum neue genomische Sequenzen hinzugekommen sind. Dagegen gibt es im murinen Datensatz der VBASE2-Klasse 1 deutlich mehr Sequenzen, die nur in VBASE2 vorkommen. Die Tabelle 2.6 zeigt aber vor allem, dass eine große Stärke von VBASE2 in der Präsentation der Klasse-2-V-Gene liegt. Die bisher veröffentlichten Annotationen der Immunglobulinloci sind auf funktionelle V-Gene fokussiert. Der für VBASE2 verwendete automatische

Prozess führt dagegen eine systematische Sequenzanalyse durch und detektiert neben den Pseudogenen, die den funktionellen V-Genen noch zu mindestens siebzig Prozent ähnlich sind, auch zahlreiche V-Gen-Relikte, die sich nicht mehr eindeutig einer V-Gen-Familie zuordnen lassen. Diese systematische Sequenzanalyse bietet eine geeignete Grundlage für Untersuchungen zur Evolution der Immunglobulinloci. Von Bedeutung ist in diesem Zusammenhang auch die automatische Identifizierung von V-Gen-Orphans, also V-Gen-ähnlichen Sequenzen, die außerhalb der Immunglobulinloci im Genom vorkommen. So wurde durch den V-Gen-Analyse-Prozess beispielsweise ein bisher nicht beschriebenes murines IgH-V-Orphan auf Chromosom 8 detektiert, musIGHV370. Dieses Segment weist vor allem in der FR3-Region große Ähnlichkeit zu V-Segmenten der V_hJ558-Familie auf. Eine Untersuchung der genomischen Umgebung des Orphans auf Chromosom 8 könnte Hinweise darauf geben, wie die Translokation des Segments stattgefunden hat.

Zusammenfassend lässt sich feststellen, dass VBASE2 zahlreiche neue Keimbahn-V-Gene von Maus und Mensch präsentiert und eine umfassende Sammlung an Pseudo-V-Genen und V-Gen-Relikten enthält. Es fehlen aber in VBASE2 auch V-Gene, die in anderen Datenbanken als Keimbahn-V-Gene ausgewiesen sind, so dass Bedarf an der Weiterentwicklung von VBASE2 besteht.

3.2.3 Erweiterungsmöglichkeiten von VBASE2

Es gibt verschiedene Ansatzpunkte, den aktuellen V-Gen-Analyse-Prozess im Sinne einer erweiterten V-Gen-Annotation zu verbessern. So sollen in Zukunft die RSS-Elemente am 3'-Ende von V-Genen im VBASE2-Eintrag angezeigt werden. RSS-Elemente werden im V-Gen-Analyse-Prozess bereits detektiert, so dass nur noch die Integration dieser Information in den Prozess der Datenbankerzeugung erforderlich ist. Eine Erweiterung der Annotation ist außerdem am 5'-Ende der V-Gene geplant: Bisher analysiert der Prozess die V-Gene nur anhand der Sequenzen, die die Kette eines reifen Antikörpers kodieren. Die Sequenz des Signalpeptids wird nicht betrachtet, da für diesen Bereich noch keine DNAPLOT-Master-Sequenzen existieren. Durch Erstellung der erforderlichen Master-Sequenzen kann das Signalpeptid und Exon 1 in zukünftigen Versionen der Datenbank auch annotiert werden. Im Zusammenhang mit der geplanten Annotation von Exon 1 der V-Gene im IgH-Locus der 129/Sv-Maus (siehe Kapitel 3.6.2) sollen Methoden entwickelt werden, die die einfache Extraktion dieser Master-Sequenzen aus den öffentlichen Datenbanken ermöglichen. Auch die in Kapitel 3.6.2 beschriebene geplante Untersuchung der Promotorsequenzen kann möglicherweise als Grundlage für eine Erweiterung von VBASE2 genutzt werden: wenn ein automatisiertes Verfahren zur Analyse der Promotorsequenzen zur Verfügung stünde, könnte dies selbstverständlich in den Datenbank-Generationsprozess integriert werden und Transkriptionsfaktorbindestellen (TFBS) können im VBASE2-Eintrag angezeigt werden. Dabei wäre es optimal, Verweise auf die eukaryotischen Promotor-Datenbanken EPD (Eucaryotic Promoter Database; http://www.epd.isb-sib.ch/) oder TRANSFAC (http://www.gene-regulation.com/).anbieten zu können. Es bleibt jedoch zu prüfen, inwieweit die Automatisierung der Promotoranalyse möglich ist und zu qualitativ akzeptablen

Ergebnissen führt.

In der aktuellen VBASE2-Version werden V-Gen-Allele nicht gekennzeichnet, da eine sequenzbasierte Unterscheidung zwischen zwei ähnlichen V-Genen desselben Haplotyps und zwei V-Gen-Allelen nicht möglich ist. Für eine zuverlässige Allelenzuordnung ist die Kenntnis aller V-Segmente der beiden beteiligten Haplotypen erforderlich, und diese Voraussetzung ist nur für wenige Einzelfälle gewährleistet. Es ist aber denkbar, zumindest den Haplotyp und den Mausstamm in den VBASE2-Datensatz mit aufzunehmen, sofern diese im EMBL-Eintrag vermerkt sind. Neben der Einführung einer solchen Haplotypen- und Stammzuordnung ist auch die Zuordnung potentieller somatischer Mutationen möglich. Alle V(D)J-Rearrangements ohne passendes Keimbahn-V-Gen können dem ähnlichsten Keimbahn-V-Gen zugeordnet werden, und das Ergebnis kann in Form einer neuen VBASE2-Klasse für potentielle somatische Mutationen in die Datenbank eingehen. Der automatische Prozess ist dabei für die Sortierung der enormen Zahl von Rearrangements sehr nützlich. Der Zusammenhang zwischen einer Keimbahn- und einer möglicherweise mutierten V-Gen-Sequenz ist aber rein hypothetisch, und eine VBASE2-Klasse mit allen bekannten Rearrangements hätte wahrscheinlich einen Umfang von mehreren Tausend V-Genen pro Locus. Die Etablierung dieser potentiellen VBASE2-Klasse 4 könnte allerdings auch für eine Erweiterung der VBASE2-Klasse 1 genutzt werden: Für Klasse-1-Sequenzen ist bisher eine hunderprozentige Übereinstimmung des Rearrangements mit der Keimbahn-konfigurierten Sequenz erforderlich, so dass V-Segmente, deren Rearrangement bisher nur in somatisch mutierter Form bekannt ist, nicht erfasst werden. Um diese Lücke zu schließen, könnte die VBASE2-Klasse 1* hinzugefügt werden, welche Klasse-2-Sequenzen enthält, die einen sehr guten Treffer (zum Beispiel 95%) in Klasse 4 aufweisen. Darüber hinaus könnte auf Grundlage des V-Gen-Analyse-Prozesses auch eine

gesonderte Datenbank für CDR3-Regionen angelegt werden, in der alle Rearrangements der EMBL-Bank beispielsweise durch die beteiligten J-Segmente klassifiziert werden könnten.

Unabhängig von möglichen Erweiterungen der VBASE2-Klassen-Struktur bietet die automatische Erzeugung von VBASE2 den Vorteil, dass der Datensatz durch Verwendung zusätzlicher Input-Sequenzen und entsprechender Modifikation des Prozesses erweitert werden kann. Voraussetzung ist die Verfügbarkeit von Master-Sequenzen der entsprechenden V-Gene, D- und J-Segmente, die als Vorlagen bei der automatischen V-Gen-Analyse dienen. Auf diese Art sollen in einer zukünftigen Version von VBASE2 die T-Zell-Rezeptor-V-Gene von Maus und Mensch, die eine ähnliche Struktur aufweisen wie die Immunglobulinloci, in den Datensatz einbezogen werden. Die T-Zell-Rezeptorloci sind zwar bereits recht gut untersucht [Lefranc und Lefranc, 2004]. VBASE2 bietet jedoch durch den DAS-Server die Möglichkeit der Darstellung der V-Gene in Ensembl und klassifiziert den aktuellen EMBL-Bank-Datensatz. Außerdem kann der Analyse-Prozess dazu genutzt werden, die Immunglobulinloci von Ratte und Schimpanse zu annotieren. Die erste Version des Schimpansen-Genoms wurde kürzlich veröffentlicht [Tarjei et al., 2005]. Die Genome beider Organismen sind in Ensembl enthalten, und ihre Immunglobulinloci sollten eine große Ähnlichkeit zu denen von Maus und Mensch aufweisen [Bruggemann et al., 1986; Meek et al., 1991; Dammers und Kroese, 2001]. Daher könnten die murinen und humanen V-, D- und J-Segmente dazu benutzt werden, die Lokalisation der entsprechenden Segmente in Ratte und Schimpanse festzustellen, soweit die Sequenzen noch nicht bekannt sind. Nach Erstellung der nötigen Master-Sequenzen können die V-Gene von Ratte und Schimpanse im VBASE2-Generationsprozess analysiert und das Ergebnis in der Datenbank dargestellt werden. Die Immunglobulin- und T-Zell-Rezeptorloci von Ratte und Schimpanse

sind bisher relativ wenig untersucht worden, und beide Spezies stellen als enge Verwandte von Maus und Mensch geeignete Ziele für die Untersuchung der Evolution der Immunglobuline dar.

Die Einbeziehung der V-Gene anderer Modellorganismen gestaltet sich schwieriger, da bei Kaninchen, Huhn und den meisten anderen Nutztieren die Genkonversion und/oder intensive somatische Hypermutation als Diversifikationsmechanismus genutzt wird und die kombinatorische Diversifikation kaum eine Rolle spielt [Review: Flajnik, 2002]. Es ist aber durchaus denkbar, auch für diese Spezies ein automatisches Verfahren zur V-Gen-Analyse zu entwickeln, wobei der hier entwickelte Prozess eine Grundlage dafür zur Verfügung stellt.

Mit der ständig wachsenden Zahl an biologischen Datenbanken und Informationssystemen steigt die Notwendigkeit, diese untereinander zu vernetzen. So ist für VBASE2 die Verbindung zu Ensembl ein wichtiger Aspekt, der im Rahmen der geplanten Kooperationen mit den Annotations-Arbeitsgruppen von Vega und Ensembl noch vertieft werden soll. Auch die Verbindung zum EBI, dessen EMBL-Bank-V-Gen-Einträge bereits auf VBASE2 verweisen, ist ausbaufähig: Optimal wäre eine Verbindung von VBASE2 zum SRS-System, das den Zugriff auf zahlreiche Datenbanken nach definierten Suchkriterien ermöglicht. Daher ist es Ziel, die Kooperation mit dem EBI diesbezüglich zu erweitern. Außerdem ist es denkbar, eine direkte Verbindung zu Strukturdatenbanken wie der PDB (Protein Data Bank) herzustellen. Für diesen Zweck können die EMBL-Bank-Einträge der V-Gene nach Verweisen auf Strukturdatenbanken durchsucht werden, und ein Verweis auf eine Strukturdatenbank könnte gegebenenfalls in den VBASE2-Eintrag aufgenommen werden.

Für den Nutzer von Immunglobulindatenbanken wäre es höchst wünschenswert, wenn die unterschiedlichen Keimbahn-V-Gen-Datensätze der einzelnen Datenbanken in einer einzigen Datenbank vereint würden. In großem Stil fand eine solche Vereinigung bei den Proteindatenbanken statt, bei der mit der Gründung von UniProt im Dezember 2003 die Datenbanken PIR, TrEMBL und SwissProt zur UniProt Kowledgebase fusionierten und die Daten zahlreicher anderer Proteindatenbanken im UniProt Archive zusammengeführt wurden [Apweiler et al., 2004]. Die dabei verwendete hierarchische Struktur erfüllt übrigens einen ähnlichen Zweck wie die Klasseneinteilung in VBASE2. Eine derartige Fusion auf Ebene der Immunglobulindatenbanken wäre zwar optimal, ist aber leider in nächster Zukunft nicht zu erwarten. Es ist jedoch möglich, die Keimbahn-V-Gene der anderen Datenbanken in das Ergebnis des VBASE2-Generationsprozesses mit einzubeziehen. So könnte VBASE2 neben den V-Genen, die durch automatische Analyse detektiert wurden, alle Keimbahn-V-Gene der anderen Datenbanken anbieten und auf die entsprechende Quelle verweisen.

Darüber hinaus ist zu überlegen, ob es einen eigenen manuell annotierten Teil von VBASE2 geben sollte, der – wie die zusätzlichen Keimbahn-V-Gene – der Datenbank nach jedem Analyse-Prozess hinzugefügt wird. So könnten die V-Gene der Klasse 2 und 3 sukzessive visuell untersucht und zuverlässige Keimbahn-Sequenzen dieser Klassen als solche markiert werden. Auch examplarisch untersuchte Promotorsequenzen können auf diese Art in den Datensatz eingehen. Die Kombination aus manueller und automatischer Annotation ermöglicht die Präsentation eines sorgfältig gepflegten Datensatzes, der durch den rein automatisch erzeugten Teil jederzeit an die Dynamik der öffentlichen Datenbanken angepasst werden kann.

Abschließend läßt sich zur VBASE2-Datenbank also feststellen, dass durch den automatischen V-Gen-Analyse-Prozess eine breite Datenbasis für die Immunglobulinloci von Maus und Mensch geschaffen wird und es zahlreiche Erweiterungsmöglichkeiten gibt, die im Rahmen dieser Arbeit nur diskutiert werden können. Die Funktionalität und der Nutzen des beschriebenen V-Gen-Analyse-Prozesses zeigte sich deutlich bei der im Folgenden dargestellten Annotation des IgH-Locus des Mausstammes 129/Sv, die sich ohne diese Grundlage deutlich aufwändiger gestaltet hätte.

3.3 Die Annotation des IgH-Locus von 129/Sv

Die genomische Sequenz ist die Grundlage für alle Mechanismen, die für die Funktionalität der Immunglobulinloci und der damit verbundenen B-Zell-Entwicklung erforderlich sind. Dabei finden einerseits Rekombinations- und Mutationsereignisse in der Individualentwicklung statt, die die Spezifität des Antigenrezeptors und die Funktionalität der B-Zelle bestimmen. Andererseits findet ein Evolutionsmechanismus statt, der im Verlauf der Wirbeltier-Evolution zu unterschiedlichen Ausprägungen der Immunglobulinloci geführt hat und eine fortlaufende Diversifizierung der Loci bewirkt, die auch die Ausbildung verschiedener Haplotypen innerhalb einer Art zur Folge hat. Die Sequenzierung und Annotation der Immunglobulinloci leistet ein Beitrag sowohl zur Aufklärung der evolutionären Vorgänge als auch zur Aufklärung der B-Zell-spezifischen Diversifizierungs-mechanismen. Voraussetzung dafür ist eine hohe Abdeckung und damit hohe Qualität der Sequenz, da die Immunglobulinloci aus sich wiederholenden Sequenzabschnitten bestehen, die sich teilweise nur durch wenige Nukleotide unterscheiden.

Die in Kapitel 2.2.3 bis 2.3.6 dargestellte Untersuchung des proximalen Teils des IgH-Locus der 129/Sv-Maus basiert auf dem Vergleich bekannter Sequenzen mit der zu untersuchenden Sequenz. Die variable Region wurde größtenteils mit Hilfe des automatischen V-Gen-Analyse-Prozesses annotiert, und nur die RSS-Elemente und das 5'-Ende des zweiten Exons wurden manuell hinzugefügt. Die kodierenden Segmente der D-, J- und C-Region wurden dagegen manuell annotiert. Oft sind dabei Informationen aus der Annotation der Originalsequenzen übertragen worden, die *in silico* nicht unmittelbar überprüfbar sind. Deshalb wird in der Annotationstabelle in Anhang I.2 auf die Originalsequenzen verwiesen, um die Herkunft der Informationen zu

dokumentieren.

Zur Benennung der V-Segmente wurden die VBASE2-IDs verwendet, und dies soll nun ausführlich begründet werden: Die Nomenklatur der murinen V-Segmente ist in der Literatur sehr heterogen. Marie-Paul Lefranc hat im Zusammenhang mit der Schaffung der IMGT-Ontologie systematische Regeln für die Nomenklatur von V-Genen aufgestellt [Lefranc et al., 1999; Lefranc et al., 2004] und alle am IMGT annotierten V-Gene danach benannt. Dabei werden die Segmente nach ihren Familien benannt und die V-Gen-Familien von 1 bis n durchnumeriert. Im Fall der humanen V-Gene entspricht die IMGT-Nomenklatur den Veröffentlichungen der beiden Arbeitsgruppen, die in den neunziger Jahren die Sequenzierung und Annotation des humanen IgH-Locus vorangetrieben haben [Cook und Tomlinson, 1995; Matsuda et al., 1998]. Sie wurde mit leichten Anpassungen 1999 vom Nomenklatur-Kommittee des Humanen Genomprojektes übernommen [HUGO Gene Nomenclature Committee, http://www.gene.ucl.ac.uk/nomenclature/]. Die Igh-V-Gen-Familien der Maus wurden dagegen in der Original-Literatur meist nicht mit Nummern, sondern mit Klon-Namen und anderen Trivialnamen belegt [Kofler et al., 1992; Mainville et al., 1996; Riblet, 2004]. Das Maus-Genom-Nomenklatur-Kommittee (Mouse Genomic Nomenclature Committee, MGNC) verwendet in Abweichung von der IMGT-Nomenklatur diese Namen für die V-Gen-Familien (http://www.informatics.jax.org/mgihome/nomen/). Allerdings sind durch das MGNC bisher nur einzelne Familien-Repräsentanten benannt worden, so dass keine Regeln zur Benennung der V-Gene innerhalb einer Familie zu erkennen sind. Riblet, dessen Familien-Nomenklatur im Wesentlichen mit der des MGNC übereinstimmt, nummeriert die V-Gene von C57BL/6 innerhalb der Familien entsprechend ihrer genomischen Position [Riblet, 2004]. De Bono dagegen, der ebenfalls 2004 eine Analyse der IgH-V-Segmente in C57BL/6 veröffentlichte,

verwendet die Familiennummierierung nach IMGT, die einzelnen V-Gene nummeriert er im Gegensatz zu Riblet nach ihrer absoluten Position im Genom [deBono et al., 2004].

In dieser Arbeit wurden die Familiennamen des MGNC verwendet, weil das MGNC für die Schaffung einer einheitlichen Nomenklatur zuständig ist. Die Nummerierung der V-Gene wurde unterlassen, da, abgesehen von der Unklarkeit der dafür gültigen Regeln, die Sequenz noch Lücken aufweist, die eventuell V-Gene enthalten können. Aus diesem Grund ist keine zuverlässige Numerierung der V-Gene möglich. Weil aber irgendeine Form der Benennung der Gene für die Annotation erforderlich ist, wurden die VBASE2-IDs eingesetzt. Die Nummer in der VBASE2-ID folgt keiner Systematik und gibt keinen Hinweis auf die Familie. Sie ermöglicht aber das einfache Nachschlagen der Namen von Riblet und de Bono und des IMGTs und eventueller weiterer Namen des V-Gens in VBASE2. Dies soll jedoch keinesfalls die Einführung einer weiteren Nomenklatur darstellen, sondern ist nur eine Übergangslösung, um die Schaffung offizieller einheitlicher Nomenklatur-Regeln durch das MGNC abzuwarten.

Die Einteilung der V-Segmente in die VBASE2-Klassen 1 und 2 umgeht die Verwendung des Funktionalitätsbegriffs. Der Nachweis der Funktionalität ist gegeben, wenn ein V-Gen in einer Proteinsequenz oder in der mRNA einer Antikörper-produzierenden B-Zelle nachgewiesen wird. Der automatische Analyse-Prozess weist jedoch nur nach, ob ein V-Segment zum Rearrangement befähigt ist. Um die Funktionalität der V-Segmente zu überprüfen ist daher eine visuelle Betrachtung der EMBL-Bank-Einträge der zugeordneten Rearrangements nötig, um festzustellen, ob es sich dabei um entsprechende mRNAs handelt.

In der variablen Region von 129/Sv wurden fünfzehn V-Segmente annotiert, die keine offensichtlichen Defekte aufweisen, aber nicht in einem Rearrangement identifiziert werden konnten. Bisher wurde allerdings nur Exon 2, welches das V-Segment enthält, und das folgende RSS-Element untersucht. Die besagten V-Segmente tragen also möglicherweise Mutationen im Exon 1, an der Spleißstelle oder in der Promotorregion, die bisher nicht untersucht wurde. Das Segment musIGHV402 könnte Repräsentant einer bislang unbekannten V-Gen-Familie sein. Dies kann jedoch nur durch Detektion eines entsprechenden Rearrangements bewiesen werden. Wahrscheinlicher ist, dass ein Defekt im 5'-Bereich des V-Gens, im RSS Spacer oder in der Proteinstruktur das Rearrangement beziehungsweise die Funktionalität verhindert und musIGHV402 als V-Gen-Relikt einzuordnen ist. Die Tabelle 2.15 zeigt, dass musIGHV402 Ähnlichkeit hat mit musIGHV128, einem Klasse-1-V-Gen der V_h36-60-Familie.

Die RSS-Elemente wurden nach den Konsensus-Regeln annotiert, die durch systematische Untersuchung der Rekombinationseffizienz von Heptamer- und Nonamersequenzen aufgestellt wurden [Review: Feeney et al., 2000]. Dabei ist aber die Position des Nonamers und die damit verbundene Spacer-Länge durch Anwendung der Regeln nicht immer eindeutig zu bestimmen. Hier wurde bei der V_h-GAM3.8, V_h-SM7- und V_h-15-Familie ein 22 bp Spacer annotiert, da das Nonamer dadurch in den Positionen 5 bis 9 dem Konsensus entspricht und die wichtige Position 7 ein Adeninnukleotid erhält. Interessant ist in diesem Zusammenhang auch die Feststellung von de Bono, alle V_hQ52-RSS von C57BL/6 hätten einen 22 bp Spacer [de Bono et al., 2004]. In dieser Arbeit wurden die V_hQ52-RSS mit 23 bp Spacer annotiert, da das Alignment der Elemente dies nahelegt [Abbildung 2.21]. Die V_hQ52-Segmente und angrenzenden Sequenzen von 129/Sv und C57BL/6 sind teilweise identisch, so

dass es sich hier offensichtlich um unterschiedliche Interpretationen derselben Sequenz handelt. Um eine weitere Interpretation zu betrachten, könnte man eine statistische Auswertung der Sequenz nach Cowell vornehmen [Cowell et al., 2002]. Diese Gruppe verwendet statistische Methoden zur Analyse von regulatorischer DNA und veröffentlichte ein Programm, das RSS-Elemente in der DNA erkennt. Grundlage dieses Programms sind allerdings RSS-Sequenzen, die teilweise widerum auf theoretischer Annotation beruhen. Eine definitive Aussage über die Sequenz und Position des Nonamers kann nur auf experimenteller Ebene durch Untersuchung des fraglichen RSS-Elements selbst getroffen werden.

Bei der Annotation der D-Segmente wurde ein interessanter Aspekt der DST4-Familie deutlich: Das Segment DST4 unterscheidet sich von DST4.2 und DST4.3 durch die Deletion eines Nukleotids in der zweiten Position (Abbildung 2.17). DST4 wurde bereits 1993 von Feeney und Riblet entdeckt [Feeney und Riblet, 1993], während DST4.2 erst 2004 in C57BL/6 als funktionelles D-Segment annotiert wurde, obwohl es nicht in einem Rearrangement nachgewiesen werden konnte [Ye, 2004]. Die Funktionalität von DST4.3 wurde noch nicht überprüft. Tatsache ist, dass die Deletion in DST4 beziehungsweise die Insertion in DST4.2 und DST4.3 ein unterschiedliches Leseraster der Segmente zur Folge hat. Das Dμ-Protein, das nach dem DJ-Rearrangement durch ein Transkript vom Promotor des D-Segments aus gebildet wird, unterdrückt das anschließende V-DJ-Rearrangement, so dass das Raster des D-Segments selektioniert wird. Das Start-Codon der Dμ-Sequenz liegt bei allen drei Mitgliedern der DST-Familie 71 Nukleotide vor dem Beginn des D-Segments. Es stellt sich die Frage, ob DST4.2 und/oder DST4.3 deshalb in einem anderen Leseraster als DST4 kodieren, oder ob diese Segmente überhaupt funktionell sind. Eine eingehende Untersuchung der CDR3-Regionen

von Rearrangements des 129/Sv- bzw. C57BL/6-Stammes in Hinblick auf das Vorkommen und das Leseraster von DST4-Segmenten kann diese Frage klären.

Aufgrund des begrenzten zeitlichen Rahmens dieser Arbeit wurden zunächst nur die wichtigsten Merkmale der Sequenz annotiert. So steht die Annotation des Exons 1 aller V-Segmente, das den größten Teil des Signalpeptids kodiert, ebenso aus wie die Untersuchung der Promotorregionen auf Transkriptionsfaktorbindestellen. Weiterhin sind in der D- und C-Region Pseudogene zu erwarten, die in anderen Veröffentlichungen annotiert wurden [Ye, 2004; Akahori und Kurosawa, 1997]. Des weiteren ist die Sequenz noch nicht nach kodierenden Sequenzen von Nicht-Immunglobulinen durchsucht worden. Es muss auch betont werden, dass es keine experimentell ermittelten Informationen über die Größe der Lücken gibt. Daher ist nicht auszuschließen, dass durch die Aufklärung der fehlenden Sequenz weitere kodierende Segmente entdeckt werden.

3.4. Die genomische Struktur des murinen IgH-Locus

3.4.1 Repetitive Elemente

Bereits Herring et al. stellten bei der Erzeugung einer YAC-Karte des murinen IgH-Locus fest, dass der Gehalt an LINE1-Elementen ungewöhnlich hoch ist [Herring et al., 1998]. Die Repeatmasker-Analyse der hier untersuchten Sequenz vom 129/Sv-Stamm bestätigte einen etwa doppelt so hohen LINE1-Anteil der variablen Region im Vergleich zum durchschnittlichen Anteil am Genom. Ebenfalls erhöht ist der Anteil an LTR-Elementen, dagegen ist der SINE-Anteil ungewöhnlich niedrig. Es stellt sich die Frage, welche Ursache diese spezielle Verteilung von repetitiven Elementen haben könnte. Im Zusammenhang mit dem großen Anteil dieser evolutionär sehr erfolgreichen Sequenzen am Säuger-Genom gibt es Untersuchungen, die eine funktionelle Bedeutung der repetitiven Elemente vorschlagen. So entdeckten Han und Mitarbeiter, dass LINE1-Elemente die Elongantion der Transkription inhibieren und diskutieren LINE1-Elemente als allgemeinen Regelwiderstand der betroffenen Gene [Han et al., 2004]. Die Transkription spielt bei der VDJ-Rekombination eine wichtige Rolle, so dass man spekulieren kann, ob der hohe LINE1-Gehalt einen von vielen Mechanismen zur Unterdrückung der Rekombination in Nicht-B-/T-Zellen impliziert. Auch der geringe Anteil an SINEs lässt sich in diese Annahme einfügen: SINEs werden als Reaktion auf Zellstress aktiviert [Review: Schmid, 1998]. Eine Aktivierung der Transkription an den Immunglobulinloci könnte zum unkontrollierten Rearrangement in B-Zellen und Nicht-B-Zellen führen, denn zahlreiche Studien belegen einen engen Zusammenhang zwischen der Bildung steriler V-Gen-Transkripte und der Öffnung der Loci für die V(D)J-Rekombination [Reviews: Corcoran, 2005; Bassing et al., 2002]. Wenn man allerdings eine Rolle der LINE1-Elemente in der Unterdrückung der Rekombination durch Inhibierung der

Transkriptionselongation annehmen möchte, so ist zu bemerken, dass dies wohl weniger für die Unterdrückung eines unkontrollierten Klassenwechsels gilt. Die Bildung steriler Transkripte ausgehend vom Promotor des I-Exons gilt zwar als Voraussetzung für den Klassenwechsel [Review: Chaudhuri und Alt, 2004], der Gehalt an LINE1-DNA in der konstanten Region entspricht aber dem durchschnittlichen Anteil im Genom.

Ein weiterer möglicher Zusammenhang zwischen Retrotransposons und den Immunglobulinloci ist die Tatsache, dass Retrotranspositionsereignisse auch zur Übertragung von DNA-Sequenzen führen können, die stromabwärts des transponiblen Elements liegen [Review: Kazazian, 2004]. Die Vervielfältigung von transponiblen Elementen innerhalb der Immunglobulinloci stellt einen Mechanismus dar, der für die Segment-Duplikation im Bereich der V- und D-Segmente mit verantwortlich sein könnte.

Unabhängig von einer möglichen Funktion der LINE1-Elemente im IgH-Locus bergen sie die Chance, durch die Analyse ihrer Sequenzen Auskunft über die jüngere Evolution des Locus zu bekommen. LINE1-Elemente evolvieren in Schüben von Vermehrung mit anschließender Diversifizierung, so dass sich eine ausgeprägte Speziesspezifität der LINE1-Linien feststellen läßt [Silver, 1995]. Insofern eignen sich LINE1-Elemente auch als Zielsequenzen zum Vergleich der IgH-Loci verschiedener Mausstämme. Der Pip (Percent identity plot) der D-Regionen von 129/Sv und C57BL/6 zeigt beispielsweise ein LINE1-Element in der D-Region von 129/Sv, das in C57BL/6 offensichtlich fehlt (Abbildung 2.17). Abgesehen von dieser Ausnahme zeigt der Pip, dass Bereiche mit repetitiven Elementen zwischen beiden Stämmen fast vollständig konserviert sind. Im Gegensatz dazu gibt es in den Bereichen, die die

kodierenden D-Segmente enthalten, deutlich stärkere Abweichungen. Die D-Region scheint also eine konservierte Rahmenstruktur aus repeptitiven Elementen zu enthalten, die die schneller evolvierenden kodierenden Bereiche umschließt.

3.4.2 Interne Sequenzwiederholungen und V-Gen-Duplikationen

Der murine IgH-Locus enthält zahlreiche Sequenzen, die in ähnlicher Form mehrfach vorkommen und deutlich über den Bereich einzelner Gen-Segmente hinausgehen. Der Dotplot der gesamten hier untersuchten Sequenz (Abbildung 2.14) sowie der Dotplot, der die D-Region in größerer Auflösung darstellt (Abbildung 2.19), zeigen bestimmte Muster von Homologien. So lassen sich beide Dotplots in Bereiche unterschiedlich ausgeprägter interner Homologie einteilen. Bereiche, die längere homologe Abschnitte in mehrfacher Wiederholung enthalten, repräsentieren dabei bestimmte Segment-Familien. Der größte dieser Bereiche liegt im JH-proximalen Teil der variablen Region und repräsentiert den Cluster der Familien V_hQ52/V_h7183 (siehe Abbildung 2.19). Auch der V_h36-60/V_hGAM3.8-Cluster im JH-distalen Teil der hier untersuchten Sequenz zeigt sich im Dotplot durch einen Bereich mehrerer homologer Abschnitte. Stromabwärts des V_h36-60/V_hGAM3.8-Clusters liegt der längste homologe Abschnitt der Region. Ein Vergleich mit der Karte des IgH-Locus (Abbildung 2.20) zeigt, dass es sich dabei um den Bereich zwischen den drei V-Segmenten der V_hSm7-Familie musIGHV125, musIGHV150 und musIGHV141 handelt, wobei der Abschnitt zwischen musIGHV125 und musIGHV150 dem Abschnitt von musIGHV150 bis musIGHV141 ähnlich ist. In diesem Bereich befinden sich allerdings auch vier Lücken in der Sequenz, so dass nur eine vorläufige Aussage über diesen Bereich möglich ist. In der D-Region liegen alle

Segmente mit Ausnahme von DQ52 auf einem einzigem Abschnitt, der sich jeweils in unterschiedlich ausgeprägter Homologie wiederholt.

Auch die Betrachtung der Karte des IgH-Locus liefert Hinweise auf homologe Bereiche innerhalb der variablen Region. So sind innerhalb des V_hQ52/V_h7183-Clusters bestimmte Muster in Bezug auf V-Gen-Familie und -Abstand zu erkennen. Beispielsweise tritt neun mal ein V_hQ52-Segment direkt hinter einem V_h7183-Segment auf, wobei das V_h7183-Segment jeweils nicht funktionell ist. Auch den exakt duplizierten V-Segmenten musIGHV202 und musIGHV200 lässt sich ein Muster zuordnen: die beiden Kopien der Segmente schließen jeweils ein funktionelles Segment der V_hQ52-Familie ein. Diese beiden funktionellen V_hQ52-Segmente, musIGHV159 und musIGHV201, unterscheiden sich nur in einer einzigen Position in FR1. Die exakten Duplikationen sind offensichtlich jüngeren Ursprungs, da bei nicht funktionellen Segmenten eine starke Diversifizierung zu erwarten ist und die Segmente im Genom von C57BL/6 jeweils nur einmal vorhanden sind. Es ist aber nicht ausgeschlossen, dass Assemblierungsfehler dazu führen, dass die Segmente in der C57BL/6-Sequenz jeweils nur einfach vorhanden sind. Andererseits ist auch nicht mit letzter Sicherheit auszuschließen, dass zumindest die Duplikation der V-Segmente musIGHV192 und musIGHV193 in der 129/Sv-Sequenz durch einen Assemblierungsfehler zu erklären sind. Deshalb erscheint es sinnvoll, die Existenz der Duplikationen durch gezielte Sequenzierung der betroffenen Bereiche noch einmal zu verifizieren.

Das komplexe Muster an homologen Bereichen im murinen IgH-Locus bestätigt die Theorie von Ota und Nei, die für die V-Gen-Evolution einen Prozess von Duplikationen mit anschließender Diversifizierung vorgeschlagen haben (birth and death process) [Ota und Nei, 1994]. Als möglicher Mechanismus für die

Duplikation wurde bereits in Kapitel 3.4.1 die Transduktion von DNA durch LINE1-Elemente genannt. Ebenso kann homologe Rekombination zwischen V-Segmenten oder auch zwischen ähnlichen LINE1-Elementen zur Entstehung von Duplikationen beitragen. Ein ähnliches Duplikationsmuster wie am murinen IgH-Locus wurde im humanen Immunglobulin-Kappa-Locus (IgK-Locus) gefunden [Kawasaki et al., 2001]. In diesem Locus wird die Komplexität durch die invertierte Duplikation eines 360 kb-Bereichs in der variablen Region noch erhöht. Bei der Untersuchung von synonymen und nicht synonymen Nukleotidaustauschen innerhalb homologer Abschnitte ergaben sich am humanen IgK-Locus auch Hinweise auf seltene Genkonversionsereignisse. Am humanen IgH-Locus wurden derartige Hinweise bisher nicht gefunden.

Bemerkenswert ist, dass alle V-Gene und V-Gen-Relikte des murinen IgH-Locus in der kodierenden Orientierung vorliegen. Beim IgK-Locus von Mensch und Maus sind neben invertierten Pseudogenen auch funktionelle invertierte V-Segmente bekannt [Thiebe et al., 1999; Kawasaki et al., 2001], die bei der VDJ-Rekombination unter Invertierung des DNA-Bereichs zwischen dem rekombinierenden V- und J-Segment in die richtige Orientierung gebracht werden, so dass eine funktionelle kodierende Sequenz für die Kappa-Kette des Antikörpers entsteht. Dagegen kommen im IgH- und Lambda-Locus von Mensch und Maus keine invertierten V-Segmente vor. Es ist bisher nicht bekannt, warum die Invertierungen nur im Kappa-Locus auftreten.

3.5 Der Igh-Haplotyp von 129/Sv

Die Immunglobulinloci der Maus wurden 1979 von Margaret Green mit einer einheitlichen Nomenklatur versehen, und in diesem Zusammenhang wurden für die konstante Region der gebräuchlichsten Laborstämme Haplotypen benannt, darunter C57BL/6 (Igh^b) und BALB/c (Igh^a) [Green, 1979]. Ebenfalls 1979 veröffentlichte Van Snick eine Zuordnung der *Igh-1*-Region (IgG2a) von 129/Sv zum Igh^a-Haplotyp [Van Snick et al., 1979]. Die Zuordnung der Haplotyen erfolgte in beiden Fällen auf der Basis von Alloantiseren. Brodeur und Riblet untersuchten 1984 die Igh-V-Region von achtzehn Laborstämmen, darunter C57BL/6 und BALB/c, durch Southern Blot Analysen mit 24 V-Gen-Sonden. Das Ergebnis der Untersuchung war, dass sich den Haplotyen der konstanten Region im Allgemeinen die Haplotyen der variablen und D-Region zuordnen lassen. Tutter und Riblet erweiterten diese Untersuchung durch Southern Blot Analysen der variablen Region von 74 Inzucht-Mausstämmen, inklusive des 129/Sv-Stammes [Tutter und Riblet, 1988]. Der Vergleich der variablen Regionen ergab einen ausgeprägten Polymorphismus zwischen unterschiedlichen Mausstämmen. Die DNA des 129/Sv-Stammes wurde dabei unter anderem mit V-Gen-Proben der V_h7183- und V_h3609P-Familie hybridisiert. Das resultierende Bandenmuster ist identisch mit dem entsprechenden Bandenmuster von BALB/c-DNA, so dass 129/Sv dem Igh^a-Haplotyp zugeordnet wurde.

Im Rahmen dieser Arbeit wurden die D-Segmente von 129/Sv mit den D-Segmenten von BALB/c verglichen. Mit Zellen und Tieren vom Stamm BALB/c wurden und werden viele immunologische Experimente durchgeführt. Der Stamm 129/Sv wird dagegen häufig für transgene Mausexperimente genutzt, und für Untersuchungen des Immunsystems kann der Haplotyp der

Immunglobulinloci durchaus von Bedeutung sein. Daher ist es Ziel dieser und folgender Untersuchungen, den Haplotyp des IgH-Locus von 129/Sv auf Sequenzebene zu charakterisieren. Beim Vergleich der D-Segmente zeigte sich, dass die D-Region von 129/Sv sich deutlich von der BALB/c-D-Region unterscheidet. Weiterhin zeigte ein erster Vergleich der Cμ-Region einen Aminosäureaustausch zwischen 129/Sv und BALB/c zu Beginn des CH3-Exons (Daten nicht dargestellt). Diese Ergebnisse deuten an, dass die Zuordnung von 129/Sv zum Igh^{a}-Haplotyp möglicherweise revidiert oder zumindest eingeschränkt werden muss.

Um den Haplotyp von 129/Sv besser bewerten zu können, ist die Ausweitung der vergleichenden Sequenzanalyse auf die variable und konstante Region nötig. Für die Exon/Intron-Bereiche der konstanten Region von BALB/c stehen vollständige genomische Sequenzen in der EMBL-Bank zur Verfügung. Die Untersuchung der variablen Region muss sich wahrscheinlich auf den Vergleich der einzelnen V-Segmente beschränken, da für die variable Region von BALB/c bisher keine ausreichenden genomischen Sequenzen vorhanden sind. Auch die bisher erstellte genomische Sequenz von 129/Sv bedeckt weniger als die Hälfte der variablen Region, so dass über den großen Bereich der V_hJ558-Familie zunächst keine Aussage getroffen werden kann. Der Vergleich der V-Segmente erfordert zuerst eine umfangreiche Datenbank- und Literatur-Recherche der verfügbaren Keimbahn-V-Gen-Sequenzen von BALB/c, wofür die VBASE2-Datenbank eine wertvolle Grundlage bietet. Möglicherweise können dabei auch weitere V-Segmente von 129/Sv in der EMBL-Bank identifiziert werden, die im derzeit noch nicht sequenzierten Teil des Locus liegen. Trotz der zu erwartenden Lücken im JH-distalen Teil des Locus sollte ein Vergleich der bekannten V-Segmente beider Stämme ausreichen, um den mutmaßlichen Igh^{a}-Haplotyp von 129/Sv zu bestätigen oder zu widerlegen.

3.6 Ausblick

Die vorliegenden Ergebnisse bieten zahlreiche Ansatzpunkte für weiterführende Untersuchungen und Entwicklungen. Die Erweiterungsmöglichkeiten der VBASE2-Datenbank und der automatischen V-Gen-Analyse wurden bereits in Kapitel 3.2.3 erörtert. In diesem Kapitel sollen Ansatzpunkte für die Fortsetzung der Arbeit am murinen IgH-Locus diskutiert werden.

3.6.1 Vollständige Sequenzierung des IgH-Locus von 129/Sv

Für eine vollständige Charakterisierung des IgH-Locus von 129/Sv ist die Vervollständigung der genomischen Sequenz Voraussetzung. Die Größe des murinen IgH-Locus wird insgesamt auf 3 Mb geschätzt, so dass der bisher analysierte Teil etwas mehr als ein Drittel der variablen Region abdeckt [Chevillard et al., 2002]. Bei der Fortsetzung der Sequenzierung gilt es zunächst, die Lücken zwischen den vorhandenen Contigs zu schließen. Sofern dies mit den bisher angewandten Sequenziermethoden nicht möglich ist, ist die Einbeziehung anderer Methoden, beispielsweise die gezielte Subklonierung der entsprechenden Bereiche, erforderlich. Auch eine Überprüfung des Sequenzbereichs der Duplikationen von musIGHV192 und musIGHV193 ist, wie bereits erwähnt, notwendig. Für die Sequenzierung des distalen Teils stehen weitere BAC-Sequenzen zur Verfügung. Abgesehen von dem schon erwähnten BAC 4k8 wurden bis zur Fertigstellung dieser Arbeit drei weitere BACs teilweise oder vollständig sequenziert, die voraussichtlich den Bereich der V_hJ606-Familie und möglicherweise auch den Beginn der V_hJ558-Familie abdecken. Für die Sequenzierung der V_hJ558-Region sind noch größere Schwierigkeiten zu erwarten als bei der Aufklärung des V_hQ52/V_h7183-Bereichs, da vermutlich

zahlreiche Sequenzabschnitte mit ausgeprägten internen Ähnlichkeiten die Mitglieder dieser größten V-Gen-Familie des murinen IgH-Locus beherbergen. Die C57BL/6-Sequenz in Ensembl weist sowohl im V_hQ52/V_h7183-Bereich als auch im V_hJ558-Bereich Lücken auf. Die unabhängige Sequenzierung des IgH-Locus eines zweiten Mausstammes mit hoher Sequenzabdeckung ist vor dem Hintergrund der Schwierigkeiten und Unsicherheiten, die wegen der zahlreichen Sequenzwiederholungen in diesem Locus auftreten, ausgesprochen sinnvoll.

3.6.2 Eweiterung der Annotation des IgH-Locus

In Kapitel 3.3 wurde bereits angesprochen, dass die vorgenommene Annotation des IgH-Locus von 129/Sv durch den zeitlichen Rahmen dieser Arbeit begrenzt wurde, und es soll nun beschrieben werden, welche weiteren Annotations- und Analyseschritte geplant sind.

Zunächst soll das Signalpeptid-kodierene Exon 1 der V-Segmente annotiert werden. Zur Identifizierung des ersten Exons eines V-Segments können entweder Sequenzvergleiche mit bekannten V-Gen-Exon-1-Sequenzen durchgeführt werden, oder die Exon/Intron-Grenzen werden aufgrund ihrer funktionellen Eigenschaften wie beispielsweise konservierter Speißstellen detektiert. Die Anwendung beider Methoden sollte die Qualität der Annotation optimieren. Zur Ermittelung von V-Gen-Exon-1-Sequenzen kann der VBASE2-Datensatz dienen, indem die Annotation von Keimbahn-konfigurierten V-Gen-Sequenzen ausgewertet wird. Als weitere Datenbanken für annotierte murine Immunglobuline stehen SwissProt/KB und die IMGT/LIGM-Datenbank zur Verfügung. Auch Signalpeptid-Datenbanken wie SPdb (http://proline.bic.

nus.edu.sg/spdb/) können nach Immunglobulin-Signalpeptiden durchsucht werden. Ergänzend kann eine BLAST-Suche gegen EST (expressed sequence tags)-Sequenzen die Lokalisierung von Exon 1 ermöglichen. Andererseits können klassische Gen-Findungs-Programme wie GENSCAN angewendet werden, um die Position von Exon 1 zu ermitteln. GENSCAN kann zwar nicht das Ende des V-Segments detektieren, da das vollständige zweite Exon erst durch die VDJ-Rekombination gebildet wird und dementsprechend am Ende des V-Segments keine Spleiß-Donorstelle, sondern ein RSS-Element vorhanden ist. Das Exon 1 könnte aber dennoch von GENSCAN erkannt werden, da GENSCAN auch unvollständige Gene detektiert. Das Ergebnis der Annotation wird visuell auf das Vorhandensein der Speißdonor- und Speißakzeptorsequenzen geprüft.

Eine weitere wichtige Aufgabe ist die Annotation cis-regulatorischer Sequenzen im 5'-Bereich der V-Gene. Die Transkription der V-Segmente wird direkt vor dem Rearrangement aktiviert, und ein Zusammenhang zwischen der Promotorsequenz und der Rekombinationshäufigkeit eines V-Gens erscheint möglich. Man nimmt an, dass die Regulation der sterilen Transkripte im Pro-B-Zell-Stadium durch die gleichen Sequenzen vermittelt wird wie die Regulation der Transkiption funktioneller antikörperkodierender DNA. Es ist bereits eine Reihe von Sequenzmotiven bekannt, die üblicherweise stromaufwärts der Transkriptionsstartstelle von V-Segmenten vorhanden sind und mit der Aktivität des V-Gen-Promotors in Verbindung gebracht werden [Review: Max, 1999]. Neben einer TATA-Box gibt es ein konserviertes Oktamer-Motiv ATGCAAAT, das allen bisher untersuchten Promotoren von Immunglobulin-V-Genen gemeinsam ist. Darüber hinaus gibt es IgH-V-typische Motive wie das Heptamer CTCATGA, das gewöhnlich 2 bis 22 bp stromaufwärts des Oktamers liegt. Auch ein pyrimidinreiches Motiv stromaufwärts des Heptamers ist beschrieben, weiterhin ein AT-reiches Motiv im V_hS107-Promotor. Die Annotation dieser und

weiterer bekannter Motive in der vorliegenden Sequenz soll das Gesamtbild des IgH-Locus von 129/Sv vervollständigen.

Die Verfügbarkeit der genomischen Sequenz und die Fortschritte in der *in-silico*-Analyse cis-regulatorischer DNA-Sequenzen legt eine weitergehende Untersuchung der Promotorregion in Bezug auf mögliche Transkriptionsfaktor-Bindungsstellen (TFBSs) nahe. Es gibt eine Vielzahl von Programmen und Algorithmen, die die Vorhersage dieser cis-regulatorischen Sequenzen zum Ziel haben. Allgemein werden dabei zwei grundsätzlich verschiedene Ansätze verfolgt: Eine Sorte von Programmen basiert auf der Erkennung von konservierten Motiven in einer Liste von Promotorsequenzen, deren Gene als co-reguliert angenommen werden. Eine aktuelle Bewertung dieser sogenannten de-novo-Methoden vergleicht eine Auswahl der für diesen Zweck zur Verfügung stehenden Programme [Tompa et al., 2005]. Die andere Sorte von Programmen berücksichtigt bei der Analyse die Sequenzmuster bekannter TFBSs, die in Datenbanken wie TRANSFAC oder EPD (Eucaryotic Promoter Database) gesammelt werden [Matys et al., 2003; Schmid et al., 2004]. Die Schwierigkeit bei der Vorhersage von TFBSs besteht darin, dass die Sequenzmotive mit etwa 6 bis 20 bp sehr kurz und außerdem degeneriert sind, so dass bei allen Vorhersage-Programmen die Abschätzung zwischen Signal und Rauschen von großer Bedeutung ist. Es könnte deshalb sinnvoll sein, mehrere Verfahren parallel anzuwenden und die Ergebnisse auf Überschneidungen zu untersuchen. Eine Analyse der Igh-V-Promotoren sollte deshalb zumindest eine de-novo-Methode und ein Verfahren zur Mustererkennung beinhalten. Weiterhin sollte gezielt nach Bindungsstellen von Transkriptionsfaktoren gesucht werden, deren regulatorische Funktion in der frühen B-Zell-Entwicklung bekannt ist [Review: Busslinger, 2004].
Eine Untersuchung regulatorischer Elemente im IgH-Locus sollte auch Matrix-

bindende DNA-Regionen (matrix attachment regions, MARs) mit einbeziehen, da die VDJ-Rekombination wesentlich durch die Chromatinstruktur und die dadurch bedingte Zugänglichkeit der variablen Region beeinflusst wird. Lokale Veränderungen in der Chromatinstruktur wie Acetylierung der Histone oder Demethylierung der DNA werden mit MARs in Verbindung gebracht. Es ist denkbar, dass die lokale Chromatinstruktur auch Einfluss auf die Rekombinationshäufigkeit einzelner V-Segmente haben kann. In diesem Zusammenhang veröffentlichten Feeney und Mitarbeiter 2002 das Ergebnis einer experimentellen Analyse der MARs und MAR-bindendenen Proteine in den flankierenden Sequenzen von 13 V-Segmenten von Mensch und Maus, darunter V_H81x (V_h7183-Familie), V_hS107-V1 und v186.2 (V_hJ558-Familie) [Goebel et al., 2002]. Sie stellten fest, dass in der Umgebung der untersuchten V-Segmente überdurchschnittlich viele MARs und auch Bindestellen für die MAR-bindenden Proteine Cux/CDP (Cut-like Protein x/CCAAT-Displacement Protein) und Bright (B Cell regulator of IgH Transcription) vorhanden sind. Da die untersuchten V-Segmente aus verschiedenen Familien und Bereichen der variablen Region stammen, wurde postuliert, dass alle V-Segmente diese Merkmale aufweisen. Diese Hypothese könnte mit geeigneten *in-silico*-Methoden überprüft werden. Engel et al. führten eine derartige *in-silico*-Analyse in zwei J-C-Regionen des murinen Lambda-Locus durch und detektierten dabei zwei S/MARs (scaffold/matrix attachment regions), die auch experimentell verifiziert werden konnten [Engel et al., 2002]. Es kann daher angenommen werden, dass mit den von Engel verwendeten Methoden auch die Detektion potentieller MARs in der variablen Region des murinen IgH-Locus möglich ist.

Neben der Chromatinstruktur spielt auch die Lokalisation der Immunglobulinloci im Zellkern für die V(D)J-Rekombination und ebenso für die allele Exklusion eine Rolle [Roldan et al., 2005; Goldmit et al., 2005]. Bisher ist nicht bekannt,

welche DNA-Sequenzmotive die Relokalisation der Immunglobulinloci vermitteln. Es ist zu erwarten, dass derartige Sequenzmotive für die jeweiligen Loci spezifisch sind, und die Suche nach IgH-spezifischen Motiven könnte Kandidaten für eine experimentelle Untersuchung dieser Sequenzen liefern. Eine mögliche Herangehensweise zur Identifizierung locusspezifischer Sequenzmotive ist eine BLAST-Suche von Fragmenten des Locus gegen das Genom der Maus. Wenn man eine Fragmentgröße von 50 bp wählt und die doppelte, überlappende Abdeckung des IgH-Locus von C57BL/6 gewährleisten möchte, wären dafür 1,2 Millionen BLAST-Suchen nötig. Das für die Generation von VBASE2 verwendete Linux-Cluster stellt für diese Aufgabe eine geeignete Hardware zur Verfügung. Die Wahl einer geeigneten Fragmentgröße und die Überlappung der Fragmente muss in mehreren Suchdurchgängen ermittelt werden. Der Erfolg dieser Herangehensweise hängt wesentlich von der Größe und Einzigartigkeit möglicher locusspezifischer Elemente ab: Je größer die Elemente und je höher die Spezifität, um so einfacher ist die Detektion.

3.6.3 Evolution des murinen IgH-Locus

Wie bereits mehrfach erwähnt wurde, unterlagen die Immunglobulinloci im Verlauf der Evolution einer starken Diversifizierung, und der von Ota und Nei vorgeschlagene Mechanismus der V-Gen-Evolution durch Duplikation und anschließende Diversifizierung findet allgemeine Anerkennung. Die Verfügbarkeit der genomischen Sequenz des murinen IgH-Locus von verschiedenen Haplotypen ermöglicht die detailgenaue Untersuchung der evolutionären Veränderung der Sequenz bis hin zur Beschreibung einzelner Evolutionsereignisse.

Erste Sequenzvergleiche zwischen den Mausstämmen 129/Sv, BALB/c und

C57BL/6 wurden im Rahmen dieser Arbeit beim Vergleich der D-Segmente bereits durchgeführt (siehe Tabelle 2.13). Zur Untersuchung der phylogenetischen Verwandtschaft der D-Regionen in den drei Stämmen wäre die Erstellung eines sequenzbasierten Stammbaums der D-Segmente, beispielsweise mit dem Phylip-Programm, sinnvoll. Die konservierten Bereiche, die im Dotplot der D-Regionen als diagonale Linien deutlich werden (siehe Abbildung 2.19), können ebenfalls als Zielobjekte zum Vergleich der D-Regionen dienen: Die Auswertung synonymer und nicht-synonymer Austausche in diesen konservierten Abschnitten kann Auskunft über ihre Entstehungszeit durch Duplikationsereignisse geben. Darüber hinaus bietet die Gesamtstruktur der D-Region aus stärker konservierten Bereichen mit repetitiven Elementen einerseits und schneller evolvierenden Bereichen mit kodierenden Segmenten andererseits ein ideales Ziel für weiterführende Untersuchungen der Evolution der D-Region. Mit Hilfe der konservierten J-Region lassen sich BAC-Klone anderer Mausstämme und Mausarten isolieren, deren Sequenzierung Grundlage für eine detaillierte Beschreibung der evolutionären Diversifizierungsereignisse der D-Region liefern würde. In diesem Sinne könnte die Untersuchung von D-Regionen der Familie der Muridae als Modell für die Evolution der Immunglobulinloci dienen.

Die repetitiven Elemente könnten auch selbst als Objekte einer vergleichenden Analyse dienen. Neben der in Kapitel 3.4.1 bereits angesprochenen Linien-Spezifität der LINE1-Elemente in der Maus könnte auch die Gesamtstruktur und Zusammensetzung der repetitiven Elemente an den Immunglobulinloci verschiedener Spezies untersucht werden. In diesem Zusammenhang könnte die interessante Frage geklärt werden, ob der hohe LINE1-Gehalt und der niedrige SINE-Gehalt für den murinen IgH-Locus charakteristisch ist, oder ob es sich dabei möglicherweise um eine generelle Eigenschaft der

Immunglobulinloci handelt.

Eine Ausweitung der Stamm-vergleichenden Sequenzanalysen auf die variable Region kann die Ergebnisse bisheriger phylogenetischer Analysen des murinen IgH-Locus ergänzen. Die in dieser Arbeit untersuchte Sequenz enthält 27 V-Segmente der VBASE2-Klasse 1, die im VBASE2-Datensatz keine andere Keimbahn-konfigurierte Referenz als die vorliegende Sequenz haben. Das bedeutet zum einen, dass diese V-Gen-Sequenzen in der Keimbahn-Konfiguration bisher nicht bekannt waren, und zum anderen, dass es sich höchstwahrscheinlich um polymorphe Allele zu V-Genen des C57BL/6-Locus handelt. Die Zuordnung der V-Gen-Allele in C57BL/6 und 129/Sv und der anschließende Vergleich der Allelen-Paare ermöglicht die Untersuchung von Substitutionsmustern zwischen den Haplotypen. Eine solche Untersuchung wurde kürzlich zu den V-Gen-Allelen vom humanen IgH-Locus sowie zu den V-Gen-Allelen vom Kappa- und Lambda-Locus von Maus und Mensch veröffentlicht [Romo-Gonzalez und Vargas-Madrazo, 2005 (1); Romo-Gonzalez und Vargas-Madrazo, 2005 (2)]. Romo-Gonzalez und Vargas-Madrazo konnten in ihrer umfassenden Analyse zeigen, dass unterschiedliche Substitutionsmuster innerhalb der einzelnen Loci, Segmente und Subregionen der Segmente sich mit den verschiedenen Selektionsdrücken auf die jeweiligen Sequenzabschnitte in Verbindung bringen lassen. Eine entsprechende Untersuchung zum murinen IgH-Locus auf Basis der vorliegenden Daten wäre eine sinnvolle Ergänzung dieser Studien.

Auch eine phylogenetische Analyse aller bisher bekannten V-Gene vom 129/Sv-Mausstamm kann zur Aufklärung der Evolution des IgH-Locus beitragen. Ein phylogenetischer Stammbaum der V-Gen-Sequenzen bildet die Verwandtschaftsstruktur der V-Gen-Familien ab. Dabei ist auch die Position der

V-Gen-Relikte interessant. Der in Tabelle 2.15 zusammengefasste Vergleich der V-Gen-Relikte mit V-Genen der VBASE2-Klasse 1 deutet an, dass die insgesamt 45 Relikte aus nur zwölf Vorläufer-Segmenten hervorgegangen sind. Der phylogenetische Stammbaum stellt auch dar, ob sich jedem Relikt ein direktes funktionelles Vorläufer-V-Gen zuordnen lässt, oder ob Relikte auch durch Duplikation und Diversifizierung anderer Relikte oder Pseudogene entstehen können. Weiterhin zeigt der Stammbaum an, welche V-Gen-Sequenzen und -Familien sich besonders erfolgreich vermehrt haben. Die Frage, ob diese Vermehrung aufgrund von Selektion der funktionellen Sequenz geschehen ist, oder ob es sich um einen eher zufällig bedingten Prozess handelt, der die genomische Struktur und die Position der V-Segmente im Locus widerspiegelt, ist allerdings kaum zu beantworten.

Die Besonderheit der Evolution der Immunglobuline und T-Zell-Rezeptoren besteht darin, dass die funktionelle Sequenz in der Keimbahn gar nicht vorhanden ist, sondern erst auf somatischer Ebene gebildet wird. Die Existenz zahlreicher Segment-Kopien, die in Verbindung mit der somatischen Mutation der Immunglobuline möglicherweise alternativ gebraucht werden könnten, hat einen geringen Selektionsdruck auf die Sequenz eines einzelnen Segments zur Folge. Dieser geringe Selektionsdruck ermöglicht eine schnelle Diversifizierung der Segmente, die einerseits den Vorteil einer zunehmenden Antikörper-Variabilität und andererseits die Entstehung zahlreicher Pseudogene zur Folge hat.Aufgrund des fortwährenden Duplikations- und Diversifizierungsprozesses zeigen die Immunglobulin- und T-Zell-Rezeptor-Loci auch die Geschichte ihrer Evolution, so dass sie geeignete Modelle zur Untersuchung evolutionärer Prozesse darstellen.

4. Material und Methoden

4.1 Hardware

Die Arbeiten zur Entwicklung eines automatischen V-Gen-Analyse-Prozesses und zur Sequenz-Analyse wurden im Wesentlichen auf einem PC (Personal Computer) durchgeführt. Für die schnelle Ausführung des V-Gen-Analyse-Prozesses steht ein Linux Cluster mit siebzehn Knoten zur Verfügung. Jeder Knoten verfügt über zwei Prozessoren und 2 GB Arbeitsspeicher, der Festplattenspeicher beträgt 72 GB. Der VBASE2-Datenbank- und Webserver verfügt über 1 GB Arbeitsspeicher und 72 GB Festplattenspeicher.

4.2 Allgemeine Software

Tabelle 4.1: Übersicht über die verwendete Software

Software	Hersteller / Entwickler	Internet-Adresse
SuSE Linux 9.0	SuSE LINUX GmbH	http://www.novell.com/de-de/linux/suse/
Perl v5.8.1	Perl Entwickler-Gemeinschaft	http://www.perl.org/
PostgreSQL 7.3.10	PostgreSQL Global Development Group	http://www.postgresql.org/
Apache Webserver 2.0	Apache Software Foundation	http://httpd.apache.org/
PHP 4.2.2	PHP Entwickler-Gemeinschaft	http://www.php.net/
XEmacs 21	Xemacs Entwickler-Gemeinschaft	http://www.xemacs.org/

4.2.1 SuSE Linux

SuSE Linux ist ein Software-Paket, das eine Distribution des Linux-Betriebssystems und diverse weitere Anwendungen umfasst. Neben der graphischen Benutzeroberfläche KDE enthält das Paket auch Programmiersprachen wie Perl und PHP, einen Apache Webserver und Programme zur Text- und Bildverarbeitung. SuSE Linux wurde sowohl auf dem PC als auch auf dem Cluster und dem Webserver installiert. Dabei wurden neben dem Betriebssystem selbst auch zahlreiche Unix-Kommandos verwendet, die in die Skripte des automatischen V-Gen-Analyse-Prozesses integriert wurden.

4.2.2 Perl

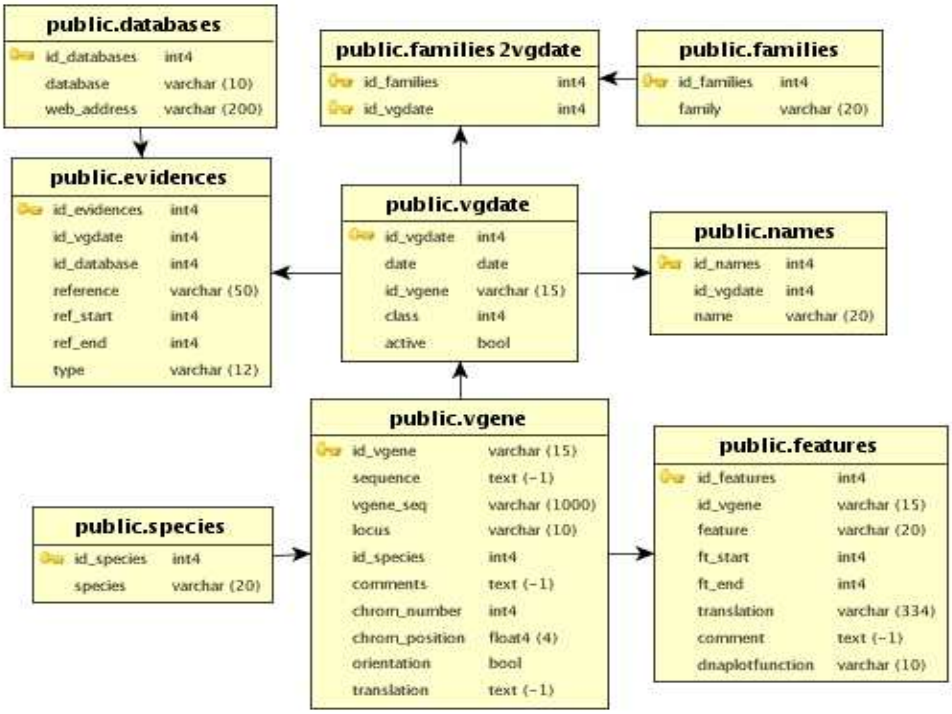
Perl ist eine freie, plattformunabhängige Programmiersprache und wird auch als Skriptsprache bezeichnet. Beim Start eines Perl-Skripts wird der Quellcode dem Perl Interpreter übergeben und direkt ausgeführt. Perl ist unter anderem bei der Entwicklung von Webanwendungen und in der Bioinformatik weit verbreitet. So gibt es unter dem Namen BioPerl eine umfangreiche Sammlung von Modulen, die speziell für die Lösung bioinformatischer Fragestellungen entwickelt wurde (http://bioperl.org/). Diese fanden in dieser Arbeit jedoch keine Anwendung, da die hier zu erfüllenden Aufgaben durch eine Kombination des DNAPLOT-Programms mit Unix-Befehlen in einfachen Perl-Skripten zufriedenstellend gelöst werden konnten. Die Perl-Skripte, die den Prozess zur automatischen V-Gen-Analyse steuern, sind in Kapitel 2.1.1 beschrieben. Der Wortlaut der Skripte kann derzeit nicht veröffentlicht werden, da die Firma Ascenion GmbH eine mögliche kommerzielle Verwertung des Generationsprozesses prüft.

4.2.3 PostGreSQL

PostgreSQL ist ein freies relationales Datenbankverwaltungssystem (RDBMS) und wurde für die Implementierung von VBASE2 genutzt. Das in Abbildung 4.1 dargestellte Schema der Datenbank wurde in engerer Kooperation mit Miguel Paulos Nunes in der Abteilung Experimentelle Immunologie der Gesellschaft für Biotechnologische Forschung in Braunschweig (GBF) entwickelt. Die Implementierung in PostgreSQL wurde von Miguel Paulos Nunes durchgeführt. VBASE2 wird bei jeder Aktualisierung vollständig neu erzeugt.

Abbildung 4.1: Schema der VBASE2-Datenbank

Das Schema wurde mit dem Programm DbVisualizer erzeugt (http://dbvis.com).



4.2.4 Apache Webserver und PHP

Der Apache Webserver ist weltweit der meist genutzte Webserver und frei verfügbar. Der Webserver dient der Informationsübertragung nach dem http-Protokoll, um eine Webseite im Internet zur Verfügung zu stellen. Die Webseite der VBASE2-Datenbank ist durch die Verwendung von PHP-Skripten dynamisch gestaltet. PHP (PHP: Hypertext Preprocessor) ist, ähnlich wie Perl, eine frei verfügbare Skriptsprache und ermöglicht die Abfrage von Daten aus der PostgreSQL-Datenbank sowie die anschließende Darstellung des Abfrageergebnisses auf der Webseite. Die PHP-Skripte für die VBASE2-Datenbank wurden von Richard Münch vom Institut für Mikrobiologie der Technischen Universität Braunschweig erstellt.

4.2.5 XEmacs

Der XEmacs ist ein frei verfügbarer Texteditor, der unter anderem zahlreiche Funktionen für die Programmierung in Perl bietet. Der XEmacs wurde in dieser Arbeit vor allem für die Erstellung der Perl-Skripte benutzt.

4.3 Bioinformatische Algorithmen und Programme

4.3.1 DNAPLOT

DNAPLOT (http://www.dnaplot.org) ist ein Programm, das für den Vergleich und die Darstellung von Immunglobulinsequenzen entwicklt wurde (Patent-Nummer DE 199 50 050 C2) [Müller et al., nicht veröffentlicht]. Das Kernstück von DNAPLOT ist ein Alignment-Algorithmus, der eine unbekannte Sequenz mit einer Vorlage-Sequenz vergleicht. Dabei werden Lücken in die unbekannte Sequenz eingefügt, bis die beste Alignment-Position der unbekannten Sequenz ermittelt ist. Für die Bewertung der Alignment-Qualität können verschiedene Scores ausgegeben werden, die die prozentuale Identität und die Alignment-Länge widerspiegeln.

Für das Alignment von Immunglobulinsequenzen sind die als Vorlagen dienenden sogenannten Master-Sequenzen von entscheidender Bedeutung. Die Master-Sequenzen sind spezies- und locusspezifisch und repräsentieren die V-Gene aller Familien eines Locus. Durch manuelles Alignment wurden sie mit Lücken versehen, die die Längenvariabilität der V-Segmente repräsentieren. Diese Lückenmuster erlauben die einheitliche Nummerierung aller V-Segmente, die für den Vergleich einzelner Segmentregionen unerlässlich ist. In dieser Arbeit wurde die IMGT-Nummerierung verwendet, bei der beispielsweise die konservierten Cysteine immer in Position 23 und 104 auftreten [Lefranc et al., 2003] Die Immunglobulin-Master-Sequenzen für Maus und Mensch wurden von Werner Müller, Leiter der Abteilung Experimentelle Immunologie, GBF, zur Verfügung gestellt und im Rahmen dieser Arbeit nicht verändert.

Neben der Alignment-Funktion hat DNAPLOT viele andere nützliche Funktionen,

die den Umgang mit DNA- und Proteinsequenzen erleichtern. Eine Auswahl ist in Tabelle 4.2 aufgestellt.

Das DNAPLOT-Programm selbst besitzt keine graphische Oberfläche, kann jedoch auf verschiedenen Webseiten zum Alignment von Immunglobulin-sequenzen genutzt werden (http://www.dnaplot.de, http://www.vbase2.org, http://vbase.mrc-cpe.cam.ac.uk, http://imgt.cines.fr). Im V-Gen-Analyse-Prozess wird das Programm aber auf dem lokalen Rechner aufgerufen.

Tabelle 4.2: Funktionen des DNAPLOT-Programms

Die Tabelle zeigt eine Aufstellung wichtiger Optionen, die in dieser Arbeit genutzt wurden.

Eingabe-Option	Funktion
align	Alignment zweier (Listen von) Sequenzen
sort	Sortierung einer Liste von Sequenzen mit der ersten Sequenz als Vorlage
comp	vergleicht zwei (Listen von) Sequenzen und gibt den Alignment-Score aus
extract	extrahiert eine Sequenz aus einer Liste von Sequenzen
first, last	betrachtet nur die angegebenen Teil-Sequenzen
fasta, gde, vbase	gibt die Sequenz(en) im angegebenen Format aus
aafmt	übersetzt DNA-Sequenz in Proteinsequenz

4.3.2 NCBI-blastall

Das blastall-Programm (http://www.ncbi.nlm.nih.gov/BLAST/) implementiert ein heuristisches Verfahren zum Sequenz-Alignment und wurde am NCBI für die schnelle Suche von Sequenzen in Datenbanken entwickelt [Altschul et al., 1997]. Die Abkürzung BLAST steht für Basic Local Alignment Search Tool. Die BLAST-Suche ist in der biologischen Forschung eine Standard-Methode für den

Zugriff auf Sequenzdatenbanken und wird auf zahlreichen Webseiten angeboten. Für die Anwendung im Rahmen der automatischen V-Gen-Analyse wurde das blastall-Programm (Version 2.2.5) aber lokal installiert und die BLAST-Suchen über die Kommandozeile ausgeführt. Dafür ist zunächst die Erzeugung einer BLAST-Datenbank aus den zu durchsuchenden Sequenzen nötig. Diese wurde durch das Programm formatdb erzeugt, welches im blastall-Programm-Paket enthalten ist.

Tabelle 4.3 zeigt eine Aufstellung der blastall-Parameter, die für den automatischen V-Gen-Analyse-Prozess angepasst wurden. Bei den Parametern, die in Tabelle 4.3 nicht genannt werden, erwies sich die Standard-Einstellung der blastall-Version 2.2.5 als geeignet. Die Einstellung des Parameters -b, die Auswahl der Input-Sequenzen sowie die Auswertung des BLAST-Ergebnisses wird in Kapitel 2.1.1.3.1 beschrieben.

Tabelle 4.3: Einstellung der Parameter des blastall-Programms

Parameter	Wert	Bedeutung
a	2	Anzahl der verwendeten Prozessoren
b	100; 5000*	Anzahl der auszugebenden Sequenzen
f	999	Schwellenwert zur Erweiterung eines Alignment-Treffers
F	F (false)	Filter der Input-Sequenzen durch das Programm DUST
m	9	Ausgabeformat (Wert 9: Tabelle)
W	20	Wortgröße

* Die Anzahl der auszugebenden Sequenzen wird in Abhängigkeit von der Größe der zu durchsuchenden Datenbank und der Anzahl der zu erwartenden Treffer gewählt. Die Standard-Einstellung im automatischen V-Gen-Analyse-Prozess ist 100; für die BLAST-Suche in den EMBL-Bank-Spezies-Datensätzen von Maus und Mensch wird -mit Ausnahme des murinen Lambda-Locus- b auf 5000 erhöht.

4.3.3 Repeatmasker

Repeatmasker (http://www.repeatmasker.org) ist eine Webseite, die ein Programm zur Detektion und Maskierung repetitiver Sequenzen in großen genomischen Sequenzen anbietet [Smit und Green, nicht veröffentlicht]. Es basiert auf einem Smith-Waterman Alignment gegen eine Datenbank mit bekannten repeptitiven Elementen sowie auf Programmen zur Detektion von Regionen geringer Komplexität. Für die Analyse der genomischen Sequenz des IgH-Locus der Mausstämme 129/Sv und C57BL/6 wurde die Version 3.0 mit den Standard-Einstellungen für die Spezies Maus benutzt.

4.3.4 PipMaker

PipMaker (http://bio.cse.psu.edu/pipmaker) ist eine Webseite, die das schnelle Alignment von zwei langen Sequenzen unter Ausgabe detaillierter Alignment-Informationen ermöglicht [Schwartz et al., 2000]. PipMaker wurde entwickelt, um konservierte Bereiche in kodierender und nicht kodierender Sequenz zu detektieren und benutzt dazu das BLASTZ-Programm, in welchem der BLAST-Algorithmus entsprechend den speziellen Anforderungen des Alignments zweier langer Sequenzen implementiert wurde [Schwartz et al., 2003]. Die wichtigste Darstellung des Alignment-Ergebnisses ist der Pip (Percent Identity Plot), bei dem die prozentuale Identität der zweiten Sequenz mit der ersten Sequenz sowie die Position ähnlicher Bereiche in einer linearen Darstellung der ersten Sequenz angegeben wird. Weiterhin kann der Sequenzvergleich als Dotplot und als Alignment mit Darstellung der einzelnen Nukleotide ausgeben werden. Wenn dem PipMaker-Programm als Input-Datei das Ergebnis einer Repeatmasker-Analyse angegeben wird, werden repeptitive Bereiche vor Durchführung des Alignments maskiert und die einzelnen repetitiven Elemente

werden anschließend im Pip und im Alignment dargestellt. Bei Angabe von Exonpositionen in der ersten Sequenz in einer zusätzlichen Input-Datei wird die Position dieser Exons in allen drei Ausgabeformen dargestellt.

Die Webseite bietet drei verschiedene Varianten des PipMakers an: PipMaker, Advanced PipMaker und MultiPipMaker. Der Advanced Pipmaker ermöglicht unter anderem die farbige Markierung von Bereichen in der Eingangssequenz, während der MultiPipMaker bis zu neunzehn verschiedene Sequenzen mit der ersten Sequenz vergleichen kann.

Im Rahmen dieser Arbeit wurden alle drei Varianten des PipMakers für zahlreiche Anwendungen genutzt. So wurde die Assemblierung der BAC-Sequenzen durch Vergleich potentiell benachbarter Sequenzfragmente fortgesetzt. Für die manuelle Annotation der konstanten Region sowie der J- und D-Segmente des IgH-Locus wurde mit dem MultiPipMaker die genomische Sequenz mit Sequenzen der EMBL-Bank verglichen, deren Einträge oder Literatur-Verweise die entsprechenden Annotationen enthalten. Die Annotation aus dem EMBL-Bank-Eintrag oder aus der Literatur wurde durch Ablesen der entsprechenden Positionen in der Alignment-Ausgabe des Programms auf die vorliegende Sequenz übertragen, wobei die EMBL-ID der jeweiligen Originalsequenz in jeder Annotation angegeben wurde. Weiterhin wurde das PipMaker-Programm zur Erstellung eines Dotplots genutzt, der interne Sequenzwiederholungen im IgH-Locus darstellt. Dazu wurde die Sequenz des 129/Sv-IgH-Locus mit sich selbst verglichen. Die 129/Sv-Sequenz wurde aber mit Hilfe des Advanced PipMakers auch mit der Sequenz des Mausstamms C57BL/6 verglichen, um konservierte und nicht konservierte Bereiche darzustellen. Die Ausgabeform Pip erlaubt dabei die einfache und übersichtliche Darstellung komplexer Informationen über die genomische

Sequenz.

4.3.5 Artemis

Artemis (http://www.sanger.ac.uk/Software/Artemis/) ist ein frei verfügbares Programm zur Darstellung und Annotation großer genomischer Sequenzen. Es wurde in dieser Arbeit im Sinne eines Sequenz-Browsers mit Annotationsmöglichkeit benutzt. Als Input-Datei diente die genomische Sequenz des murinen IgH-Locus im EMBL-Format. Annotationen, die durch den automatischen V-Gen-Analyse-Prozess oder durch PipMaker-Analysen erstellt worden waren, wurden zuvor der Sequenz im EMBL-Format angefügt. Mit Artemis wurde diese teilannotierte Sequenz visuell untersucht, bisherige Annotationen überprüft und weitere Annotationen gespeichert. So wurden die RSS-Elemente der V-, D- und J-Region durch visuellen Vergleich mit den RSS-Konsensus-Sequenzen annotiert. Weiterhin wurde der Beginn der V-Segmente so angepasst, dass jedes V-Segment mit der Spleiß-Akzeptor-Stelle von Exon 2 beginnt. Dies war nötig, weil der automatische V-Gen-Analyse-Prozess das V-Segment der reifen Polypeptidkette annotiert, so dass die Signalpeptid-kodierenden Nukleotide von Exon 2 manuell angefügt werden mussten. Die fertig annotierte Sequenz wurde im EMBL-Format gespeichert.

4.3.6 Sonstiges

4.3.6.1 GENSCAN

Das Programm GENSCAN (http://genes.mit.edu/GENSCAN) detektiert Exon/Intron-Strukturen in eukaryotischer DNA [Burge und Karlin, 1997]. GENSCAN basiert auf einem Wahrscheinlichkeitsmodell, das Spleißsignale,

Signale für die Transkription und Translation sowie die Längenverteilung und Zusammensetzung eukaryotischer Gene berücksichtigt. Der GENSCAN Webserver wurde in der vorliegenden Arbeit für die Annotation der Exons der konstanten Regionen benutzt. Dabei diente GENSCAN im Wesentlichen zur Überprüfung der aus der Literatur übernommenen Annotation. Die Anwendung auf die J-, D- und V-Segmente erbrachte keine sinnvollen Ergebnisse, da diese Segmente nicht den von GENSCAN angenommenen Eigenschaften entsprechen.

4.3.6.2 EMBOSS

EMBOSS (European Molecular Biology Open Software Suite, http://emboss.sourceforge.net/) ist ein frei verfügbares Paket von Programmen zur molekularbiologischen Sequenz-Analyse [Rice et al., 2000]. Zahlreiche Programme der EMBOSS Suite wurden bei Erstellung der vorliegenden Arbeit für kleinere Aufgaben im Umgang mit den Sequenzen benutzt, wie zum Beispiel die Programme matcher (Alignment zweier Sequenzen), revseq (reverses Komplement der Input-Sequenz), polydot (Dotplot mehrerer Sequenzen) und extractseq (extrahiert einen definierten Bereich aus einer Sequenz). Die EMBOSS Suite wurde im Rahmen des NFGN-BLAST-Servers (http://ngfnblast.gbf.de/) auf einem Linux Server der GBF installiert [Hühne et al., nicht veröffentlicht].

4.3.6.3 SeaView

SeaView (http://pbil.univ-lyon1.fr/software/seaview.html) ist ein freier graphischer Editor für multiple Sequenz-Alignments und wurde in dieser Arbeit zur einfachen visuellen Darstellung von Sequenzen im FASTA-Format benutzt.

4.3.6.4 Vector-NTI Suite

Die Vector-NTI Suite ist ein kommerzielles Paket von Programmen zur Sequenz-

Analyse auf einem Windows PC (ehemals: Firma Informax; aktuell: Firma Invitrogen, .http://www.invitrogen.com). Im Rahmen dieser Arbeit wurde für die Sequenz-Assemblierung und -Formatierung lediglich das Programm ContigExpress der Version Vector-NTI Suite 8 als Ergänzung zu PipMaker und Artemis benutzt.

4. 4 Sequenzdaten und biologische Datenbanken

4.4.1 Sequenz des IgH-Locus des Mausstammes 129/Sv

Grundlage für die Sequenzierung des IgH-Locus von 129/Sv sind 46 BAC-Klone, die Roy Riblet (Torrey Pines Institute for Molecular Studies, La Jolla, USA) der Abteilung Genom-Analyse der GBF zur Sequenzierung übergeben hat. Die Abfolge der Klone war durch eine in der Riblet-Arbeitsgruppe erstellte physikalische Karte des Locus bekannt. Bis zum April 2005 wurden 21 Klone in der Abteilung Genom-Analyse unter der Leitung von Helmut Blöcker ganz oder teilweise sequenziert, für zehn Klone konnte eine lückenlose Sequenz erstellt werden (siehe Tabelle 2.9). Die anschließende Assemblierung der BAC-Insert-Sequenzen wurde mit dem Gap4-Programm, Version 4.8b1, ebenfalls in der Abteilung Genomanalyse durchgeführt (http://staden.sourceforge.net/) [Bonfield et al., 1995]. Das Ergebnis war eine Sequenz in durchschnittlich achtfacher Abdeckung, bestehend aus 1555244 Nukleotiden, die in 32 Contigs vorlagen. Die physikalische Karte konnte dabei im Wesentlichen bestätigt werden.

Im Rahmen dieser Arbeit wurde die Assemblierung der Contigs durch Anwendung des blastall-Programms, des PipMaker-Programms und mit Hilfe von Ensembl fortgesetzt. Zur Überprüfung der Orientierung der Contigs wurde mit allen Contig-Enden eine BLAST-Suche gegen den Ensembl-Maus-Datensatz durchgeführt. Drei ungeordnete Fragmente des Klons 356c12 wurden aufgrund der Homologie zur Sequenz von C57BL/6 in eine hypothetische Reihenfolge gebracht. Anschließend wurden die Contig-Enden als Input für eine BLAST-Suche gegen die gesamte Sequenz aller 26 Contigs benutzt. Auf diese Weise wurden überlappende Regionen festgestellt. Weiterhin wurden die

vollständigen Contigs mit dem MultiPipMaker-Programm in BAC-Fragmenten oder Contigs gesucht, die auf der physikalischen Karte benachbart sind. Des weiteren wurde die Sequenz mit Hilfe der EMVEC-Datenbank auf Vektorsequenz untersucht und vorhandene Vektorsequenz aus der assemblierten Sequenz entfernt.

4.4.2 EMBL-Bank

Die EMBL-Bank (EMBL nucleotide sequence database, http://www.ebi.ac.uk/embl/) ist die primäre europäische Nukleotid-Sequenzdatenbank und wird am Europäischen Bioinformatik-Institut (EBI) in Zusammenarbeit mit den amerikanischen und japanischen Pendants GenBank und DDBJ geführt [Kanz et al., 2005]. Durch den täglichen Datenaustausch zwischen diesen drei Datenbanken wird eine Resource zur Verfügung gestellt, die praktisch alle bisher veröffentlichten Nukleotidsequenzen weltweit zugänglich macht. Die EMBL-Bank besteht aus verschiedenen Divisionen, in denen Sequenzen unter anderem nach Spezies sortiert vorliegen und nicht fertiggestellte Sequenzen aus den Genom-Sequenzierprojekten von gewöhnlichen Sequenzen getrennt werden. Der hier entwickelte automatische V-Gen-Analyse-Prozess speichert über den ftp-Server des EBI (ftp://ftp.ebi.ac.uk/) die Sequenzdatensätze von Mensch und Maus ebenso wie die Datensätze HTG (High Throughput Genomic sequences) und WGS (Whole Genome Shotgun sequences) auf dem lokalen Server. Nach Ablauf des Prozesses wird eine Textdatei erzeugt, die alle EMBL-Bank-Referenzen mit den dazugehörigen VBASE2-IDs enthält. Diese Datei wird auf einem lokalen FTP-Server im Internet zugänglich gemacht und dient der EMBL-Bank zur Erstellung der Verweise von EMBL-Bank-Einträgen auf VBASE2.

Die in dieser Arbeit untersuchte genomische Sequenz des murinen IgH-Locus ist mit der hier beschriebenen Annotation der kodierenden Sequenzen und RSS-Elemente unter der EMBL-Zugangsnummer (AC) AJ851868 zugänglich.

4.4.2.1 EMVEC

EMVEC (http://www.ebi.ac.uk/blastall/vectors.htm) ist eine Sammlung synthetischer Sequenzen aus der EMBL-Bank, die bei Klonierungsexperimenten und Sequenzierungen genutzt werden. Eine BLAST-Suche gegen EMVEC wurde genutzt, um Vektorsequenzen aus der zu assemblierenden genomischen Sequenz zu entfernen.

4.4.3 Ensembl Genom-Browser

Die Ensembl-Webseite (http://www.ensembl.org) bietet Zugang zu den Genomsequenzen von derzeit neunzehn eukaryotischen Spezies, darunter Mensch und Maus [Hubbard et al., 2005]. Der Ensembl Genom-Browser bietet dabei vielfältige Verweise und Schnittstellen zu anderen molekularbiologischen Informationssystemen und stellt die komplexen Zusammenhänge auf einer durch den Nutzer konfigurierbaren Oberfläche dar. Die Sequenzen der in Ensembl dargestellten Chromosomen von Mensch und Maus werden für den VBASE2-Generationsprozess benötigt und, ebenso wie die Sequenzen der EMBL-Bank, über einen FTP-Server lokal gespeichert (ftp://ftp.ensembl.org/pub/). V-Gene der VBASE2-Datenbank, die in den Ensembl Chromosomen enthalten sind, verweisen auf die entsprechende Sequenz-Position im Ensembl Genom-Browser. Mit Unterstützung von Andreas Kahari von der Ensembl Arbeitsgruppe am EBI wurde ein DAS (Distributed

Annotation System)-Server eingerichtet. Die Verbindung des lokalen Browsers mit dem DAS-Server ermöglicht die Darstellung der VBASE2-V-Gene im Ensembl Browser. Bei einer Aktualisierung des Ensembl-Sequenzdatensatzes muss der VBASE2-Datensatz ebenfalls aktualisiert werden, damit die korrekten Positionen der V-Gene im Ensembl Browser aufgerufen werden.

4.4.4 Weitere Datenbanken und Webdienste

In diesem Kapitel wurden die wichtigsten Webdienste und Datenbanken, die im Rahmen dieser Arbeit benutzt wurden, bereits aufgeführt. Darüber hinaus wurden zahlreiche weitere Webseiten genutzt. Ausdrücklich erwähnt werden soll die Webseite des IMGT (http://www.imgt.cines.fr), die ausgesprochen umfangreiche Informationen und zahlreiche Tools für die Untersuchung immunologischer Sequenzen anbietet. Immunlogische Datenbanken werden in Kapitel 1.2 ausführlich besprochen. Eine Liste der wichtigsten und am häufigsten genutzten Internetseiten und Webdienste befindet sich in Anhang II.4.

5. Anhang

I Ergebnisse und Tabellen

I.1 Ergebnis des TrEMBL-Filter- und Selektionsprozesses

A: Zugangsnummern der TrEMBL-Sequenzen, die durch den Filter als Immunglobuline identifiziert wurden, welche unbeabsichtigt in UniProtKB/TrEMBL aufgenommen wurden.

humane Immunglobuline:

Q8TC63, Q7Z3Y5, Q9UL87, Q7Z2E8, Q6N093, Q9NPP6, Q6ZP87, Q8WY24, Q6ZW64, Q7Z351, Q6MZX7, Q96DK0, Q96I69, Q6N097, Q6P089, Q6P055, Q6N094, O95978, Q6N095, Q6MZW0, Q6N090, Q8WVJ6, Q9H5Z4, Q9BRV0, Q9UGP3, Q8IZD8, Q9UL94, Q6N030, Q8NEJ1, Q7Z5W1, Q96QS0, Q96JD0, Q6N096, Q91Z05, Q6PIH6, Q9Y509, Q9UL90, Q8WUK3, Q6PIQ7, Q9UL88, Q6PJA3, Q9ULB6, Q7Z473, Q6GMX8, Q6GMW7, Q6P5S8, Q6GMV9, Q9UL71, Q9UL79, Q6IN99, Q6KB05, Q9UL91, Q6GMX3, Q6PJ95, Q8WUK1, Q9UL80, Q7Z2U3, Q6GMX7, Q9UL73, Q6PIK1, Q6N091, Q6N092, Q96AA6, Q6P5S3, Q6PJF1, Q7Z379, Q6PJB5, Q9UL82, Q7Z7P5, Q6GMY2, Q9Y298, Q6GMV7, Q8R3V9, Q8NEK0, Q6GMX5, Q8IZD7, Q9UL75, Q6DHW4, Q8N5K4, Q9UL77, Q8WU38, Q6PIL0, Q96EY0, Q6GMX1, Q9BU10, Q6NSA4, Q96SA9, Q7Z374, Q96KX8, O43234, Q920E7, Q6PI81, Q96JD2, Q6ZP85, Q7Z3Y4, Q6N089, Q96SB0, Q9UL70, Q6GMW6, Q9UL93, Q6PDB8, Q6GMX9, Q8N4Y9, Q6P491, Q6PIL8, Q6PJA4, Q6GMW4, Q6PJ28, Q9UL83, Q8TE63, Q9NSD6, Q6NYH3, Q6PIT5, Q6NS95, Q9UL89, Q96K68, Q6NS96, Q6IPQ0, Q6PJG0, Q96BB9, Q9UL84, Q7Z2U7, Q6MZU6, Q96PF6, Q9UL85, Q86SX2, Q6GMX4, Q9UL92, Q9UL95, Q6MZV7, Q8TCD0, Q6GMX0, Q6ZVX0, Q8N5F4, Q9UL86, Q6P4I8, Q6PJA2, Q8TBC9, Q6IN78, Q9UL96, Q9UL72, Q8TBD0, Q6GMX6, Q6P6C4, Q6GMV8, Q6PIH7, Q8N355, Q7Z3Y6, Q96E61, Q9UL81, Q6PIH4, Q6MZQ6, Q9HCC1, Q6MZX9, Q6MZV6, Q6GMX2, Q6GMW0, Q8K122, Q9BQB8, Q6PJF2, Q9UL78, Q6P2J1, Q8NCL6, Q6N041, Q6GMW1, Q9UL74, Q96JD1, O95973, Q6GMW3, Q8WUX4

murine Immunglobuline:

Q6P6C4, Q9D8L4, Q9UL93, Q811U5, Q9JL75, Q6LEM8, Q7TMK1, Q8VCX4, Q8R062, Q8TCD0, Q6PI81, Q9JL83, Q99LC4, Q6GMY2, Q8WUK1, Q6PJB2, Q6PIP8, Q6PF95, Q6P491, Q91WT1, Q8VCV5, Q91Z07, Q91XE1, Q924P6, Q8CGS1, Q7TPE3, Q8VIJ0, Q9UL80, Q924Q1, Q7TMK3, Q91WS9, Q6MZU6, Q6PKE4, Q8K1F1, Q924P5, Q924R6, Q8R3V9, Q925S3, Q8K1F0, Q924R0, Q6PJA7, Q8K172, Q924Q8, Q9ULB6, Q9JL81, Q8NEK0, Q924P8, Q920E6, Q924R2, Q99KA4, Q924Q3, Q920E7, Q920E9, Q9UL90, Q924Q5, Q99M22, Q8R3H6, Q7TMT6, Q8VDD0, Q91Z05, Q6PJA2, Q6PDB8, Q9QYF0, Q9Z1C6, Q924Q4, Q6LBQ5, Q8CGS2, Q6KB05, Q7TS98, Q9Z1C4, Q7TMK0, Q91WR1, Q9QXE9, Q8K0Z4, Q8K122, Q921A6, Q9D9B8, Q99NG4, Q9JL84, Q8VIJ1, Q91VA2, Q811C3, Q9U410, Q91V67, Q9UL91, Q924R4, Q924R8, Q925S1, Q9JL78, Q7TQM2, Q91WP5, Q9JL85, Q96BB9, Q9ET13, Q99LA6, Q7TMK4, Q924P7, Q9JL80, Q925S9, Q9JL77, Q8VCP0, Q8VDE2, Q6PJB5, Q924Q9, Q8K1F2, Q924R5, Q91X92, Q8VEA0, Q91WT3, Q80ZI7, Q920E8, Q924Q6, Q924Q7, Q9JL74, Q8K1F3, Q924R7, Q9JL79, Q6PJA4, Q9QXF0, Q8VDC9, Q924P9, Q924R1, Q924Q2, Q8K0F2, Q811U6, Q8WU38, Q924Q0, Q8VCX7, Q924R3, Q61750, Q99M11, Q9DCD9, Q9JL76, Q925S2

B: Zugangsnummern von Sequenzen der EMBL-Bank, die für die Aufnahme in UniProtKB/TrEMBL ausgewählt wurden.

humane Immunglobuline:

AF209739, AY582393, AF174081, AF103034, AY607369, M65099, AF308553, S71447, AB021511, AY393531, U84183, AB063834, U24691, AF062174, AJ244965, Z33898, AY393201, AY429916, AY607487, M34029, M18512, AY393531, Z14196, Z14189, AY429940, Z14171, U80084, AX306531, AF209739, AF062150, Z14202, AF062249, AY429778, AY190824, AF062099, X56158, AY607505, U84162, AY429898, AY055483, AF174059, AY393171, AY607542, AJ389179, AJ406714, AY607534, AJ579192, AJ239342, AJ426288, AJ426288, AY320839, AF052530, U41569, U70028, AF039298, AB064050, AJ426150, AJ415714, AF052535, AF103430, BC073764, AY043087, AF103450, AF209742, AF431054, X72422, Y17950, AF431056, AY320973, AB064099, AF103509, Z34909, AF431050, M80914, AF431056, AF386417, AB063964, AX082136, U86790, AJ556859, AF386376, AJ406816, AJ426288, AJ399849, M28078, AJ347691, AB064161, AB064177, AJ406870, AJ698335, AJ506632, AX400073, AF386212, AF073705, AB064195, X14583, AF386250, AY043124, BC020233, AJ414926, AJ698334, L29162, AJ388666, AB064184, AJ426415, AF194716, AF194585, AJ298540, U76676, AB087878, AJ698336, L03633, AJ224900, AB064196, AY043127, U86802, AF103593, AF386106, AY190821, BC018749

murine Immunglobuline:

AY172531, AY182645, M20835, M19401, AF547091, M36219, M34613, AY182676, AY089722, AJ240411, M32378, M22439, M83724, X59109, L22320, AY173966, M13281, AY172486, U62045, AF163754, AY594679, AJ229175, AY182728, AY172428, AY172457, AY090904, U42027, X55984, Y00743, AF163738, X59174, X03301, X70093, AF547077, AY182415, AY172883, AY182501, X94418, L23130, AY182741, AY182725, AF163750, U37873, L23133, X01435, AY182673, AF458160, AY172521, AF307937, U88680, AF458211, AF318457, U55484, U23094, AF318458, AF459843, M94017, AJ240410, U37862, M11700, AF045504, AF045506, AY436979, U55552, M34613, AY437060, M59921, AF006586, AF218628, M11697, AF045483, AF253061, AF021871, M17953, AF455942, AF459860, AF455340, AF455326, AF455990, U88682, AJ277813, U88677, X14623, AY436944, AB070543, M34579, AJ240415, AF045502, X12412, AY672648, AJ240294, AF218668, AJ240408, AF458162, AJ240416, U22977, AF218636, AF045501, AY547436, AY454455, X03088, AF045503, U30233, X14622, U62051, AY081858, U62050, M64153, AF021864, AF137620, M17488, M12767, X59088, AF139238, M12372, AF276279, M57586, AF045515, AY437053, M83721, X16678, M59919, X02177, AF072781, M55318, U62054, M12183, X14621, AY229950, U18586, M36758, M36261, M57588, U73592, X14097, AF137622, X58219, AY594678, X79906, M57570, U30240, X60425, X69859, AF045514, AF045512, U55631, AF045513, M34740, AB050076, Z22120, AF163743, AF045516, X70097, AF137613, X91669, J00569, M34633, M34598

I.2 Ergebnis der Annotation der V-, D-, J- und C-Region

Durch Aneinanderreihung der Contigs C02 bis C14 (Tabelle 2.10) wurde eine Gesamtsequenz erstellt, bei der die Lücken zwischen den Contigs durch 100 N-Nukleotide repräsentiert werden. Die Positionsangaben der Annotation beziehen sich auf diese Gesamtsequenz (EMBL-Bank Eintrag: AJ851868).

I.2.A Annotation der V-Region

Die V-Segmente sind durch die jeweilige ID in der VBASE2-Datenbank bezeichnet. Ein V-Segment wird in Analogie zur EMBL-Bank-Feature-Table-Definition als Bereich vom Beginn des zweiten Exons bis zum Beginn des RSS-Elements definiert.

V-Segment/ Element	Beginn	Ende	Beson-derheit	VBASE2-Klasse	Familie	Namen
musIGHV143	12948	13254	Pseudo	2	V_h36-60	-
RSS	13255	13293				
musIGHV154	50710	51014	-	1	V_h15	IGHV14S3*01
RSS	51015	51052				
musIGHV158	73257	73571	-	1	V_hS107	V11, IGHV7S3*01
RSS	73572	73610				
musIGHV157	92626	92930	-	2	V_hGAM3.8	VGK7, IGHV9S4*01
RSS	92931	92968				
musIGHV404	94378	94693	Relikt	2	-	-
RSS	94694	94731	anormal			
musIGHV401	109762	110077	Relikt	2	-	-
RSS	110078	110116				
musIGHV156	122137	122441	-	1	V_hGAM3.8	VGK1A, IGHV9S3*01
RSS	122442	122479				
musIGHV402	135707	136022	-	2	-	-
RSS	136023	136061				
musIGHV155	146585	146889	-	1	V_hGAM3.8	VGK4, IGHV9S6*01
RSS	146890	146927				
Lücke	163967	164066				
musIGHV406	167006	167318	Relikt	2	-	-
RSS	167319	167356				
musIGHV152	185158	185462	-	2	V_hGAM3.8	VGK3, IGHV9S2*01
RSS	185463	185500				
musIGHV405	190013	190328	Relikt	2	-	-
RSS	190329	190367				

V-Segment/ Element	Beginn	Ende	Beson-derheit	VBASE2-Klasse	Familie	Namen
musIGHV151	202414	202718	-	2	V_hGAM3.8	VGK1B, IGHV9S3*02
RSS	202719	202756				
musIGHV407	206933	207234	Relikt	2	-	-
musIGHV408	224573	224874	Relikt	2	-	-
musIGHV125	226520	226824	-	1	V_hSm7	IGHV14S1*01
RSS	226825	226862				
Lücke	230652	230751				
musIGHV153	238124	238430	-	1	V_h11	2C8, IGHV11S1*01
RSS	238431	238469				
musIGHV409	248735	249039	Relikt	2	-	-
musIGHV126	252909	253213	-	1	V_h36-60	IGHV3S2*01
RSS	253214	253252				
musIGHV149	272920	273226	-	1	V_hX24	VHGal55.1, IGHV4S2*01
RSS	273227	273265				
musIGHV150	288990	289294	Pseudo	2	V_hSm7	-
RSS	289295	289332				
Lücke	291847	291946				
Lücke	294845	294944				
musIGHV127	298352	298658	-	2	V_h11	-
RSS	298659	298697				
Lücke	304113	304212				
musIGHV423	310742	311049	Relikt	2	-	-
RSS	311050	311088	anormal			
musIGHV138	314921	315225	-	1	V_h36-60	-
RSS	315226	315264				
musIGHV142	334868	335174	-	1	V_hX24	IGHV4S1*01
RSS	335175	335213				
musIGHV141	356695	356999	-	1	V_hSm7	IGHV14S2*01
RSS	357000	357037				
musIGHV140	372656	372976	Pseudo	2	V_hS107	-
RSS	372977	372983	verkürzt			
musIGHV168	388411	388727	-	1	V_hS107	V1, IGHV7S1*01
RSS	388728	388766				
musIGHV424	400094	400400	Relikt	2	-	-
RSS	400401	400439				
musIGHV171	405493	405796	-	1	V_hQ52	VOx-1, IGHV2S4*01
RSS	405797	405835				

V-Segment/ Element	Beginn	Ende	Beson-derheit	VBASE2-Klasse	Familie	Namen
musIGHV170	407232	407533	Relikt	2	-	-
RSS	407534	407572	anormal			
musIGHV169	408819	409122	Relikt	2	-	-
musIGHV139	430891	431195	-	1	V_h7183	VH61-1P, IGHV5S18*01
RSS	431196	431234				
musIGHV166	455679	455985	-	1	V_h7183	-
RSS	455986	456024				
musIGHV410	487042	487352	Relikt	2	-	-
musIGHV134	494373	494676	-	1	V_hQ52	-
RSS	494677	494715				
musIGHV411	496456	496757	Relikt	2	-	-
musIGHV167	514653	514957	-	2	V_h7183	-
RSS	514958	514996	anormal			
musIGHV412	522895	523194	Relikt	2	-	-
musIGHV133	535829	536132	-	2	V_hQ52	-
RSS	536133	536171				
musIGHV174	552409	552712	-	1	V_hQ52	-
RSS	552713	552751				
musIGHV413	554488	554795	Relikt	2	-	-
musIGHV172	560414	560719	Relikt	2	-	-
musIGHV414	563447	563749	Relikt	2	-	-
musIGHV173	565095	565398	-	1	V_hQ52	-
RSS	565399	565437				
musIGHV415	570957	571255	Relikt	2	-	-
RSS	571256	571294				
musIGHV175	586945	587248	-	1	V_hQ52	-
RSS	587249	587287				
musIGHV416	589188	589493	Relikt	2	-	-
musIGHV176	595220	595523	-	1	V_h7183	VH57-1M, IGHV5S12*01
RSS	595524	595562				
musIGHV417	611268	611578	Relikt	2	-	-
musIGHV177	618120	618426	-	1	V_h7183	-
RSS	618427	618465				
musIGHV418	635018	635325	Relikt	2	-	-
musIGHV178	644005	644311	-	1	V_h7183	IGHV5S8*01
RSS	644312	644350				
musIGHV419	654809	655107	Relikt	2	-	-

V-Segment/ Element	Beginn	Ende	Beson-derheit	VBASE2-Klasse	Familie	Namen
musIGHV420	666213	666517	Relikt	2	-	-
musIGHV164	672549	672849	-	2	V_hQ52	-
RSS	672850	672888				
musIGHV165	679312	679618	-	1	V_h7183	VH68-5N, IGHV5S15*01
RSS	679619	679657				
musIGHV422	688260	688567	Relikt	2	-	-
Lücke	696602	696701				
musIGHV160	698814	699120	-	1	V_h7183	-
RSS	699121	699159				
musIGHV202	713774	714075	Relikt	2	-	Vh7183.b11
musIGHV159	715506	715809	-	1	V_hQ52	-
RSS	715810	715848				
musIGHV200	718542	718842	Relikt	2	-	VhQ52.b5
RSS	718843	718884	anormal			
musIGHV199	731327	731631	Relikt	2	-	Vh7183.b10psi
musIGHV161	739235	739541	-	2	V_h7183	VH283, IGHV5S3*01
RSS	739542	739580				
musIGHV162	747971	748274	-	1	V_hQ52	-
RSS	748275	748313				
musIGHV163	756112	756418	-	1	V_h7183	VH7183.10, IGHV5S10*01
RSS	756419	756457				
musIGHV426	762749	763053	Relikt	2	-	-
musIGHV136	764414	764717	-	2	V_hQ52	-
RSS	764718	764751	anormal			
musIGHV137	770231	770534	-	1	V_hQ52	-
RSS	770535	770573				
musIGHV427	772293	772598	Relikt	2	-	-
musIGHV122	778341	778647	-	1	V_h7183	VH98-3G, IGHV5S16*01
RSS	778648	778686				
Lücke	784042	784141				
musIGHV123	785949	786251	Relikt	2	-	-
musIGHV124	787948	788251	-	1	V_hQ52	-
RSS	788252	788290				
musIGHV431	793106	793407	Relikt	2	-	-
RSS, anormal	793408	793444				
musIGHV135	805202	805508	-	1	V_h7183	-
RSS	805509	805547				

V-Segment/ Element	Beginn	Ende	Beson-derheit	VBASE2-Klasse	Familie	Namen
musIGHV428	820143	820445	Relikt	2	-	-
musIGHV144	821827	822130	Pseudo	2	V_hQ52	-
RSS	822131	822169				
musIGHV429	837638	837949	Relikt	2	-	-
musIGHV145	845526	845832	-	2	V_h7183	VH37.1, IGHV5S4*02
RSS	845833	845871				
musIGHV146	854263	854566	-	1	V_hQ52	-
RSS	854567	854605				
musIGHV147	862404	862708	-	1	V_h7183	VH76-1BG, VH7183.9, IGHV5S9*01
RSS	862711	862749				
musIGHV432	873881	874181	Relikt	2	-	-
musIGHV132	885001	885304	-	1	V_hQ52	-
RSS	885305	885343				
musIGHV148	899546	899852	-	1	V_h7183	VH50.1, IGHV5S5*01
RSS	899853	899891				
musIGHV202	914562	914863	Relikt	2	-	Vh7183.b11
musIGHV201	916300	916603	-	1	V_hQ52	2-4b-8
RSS	916604	916642				
musIGHV200	919344	919644	Relikt	2	-	VhQ52.b5
RSS	919645	919687	anormal			
musIGHV193	932597	932901	Relikt	2	-	Vh7183.b7psi
musIGHV192	941778	942084	-	1	V_h7183	5-3b-5, Vh7183.b6, 3:3.9, IGHV5S6*01
RSS	942085	942123				
Lücke	942702	942801				
musIGHV199	952322	952626	Relikt	2	-	Vh7183.b10psi
musIGHV131	960234	960540	-	1	V_h7183	VH7183.14, IGHV5S14*01
RSS	960541	960579				
musIGHV197	966996	967298	Relikt	2	-	Vh7183.b8
musIGHV179	968700	969003	-	2	V_hQ52	-
RSS	969004	969042				
musIGHV194	974525	974825	Relikt	2	-	VhQ52.b3
RSS	974826	974864				
musIGHV193	987288	987592	Relikt	2	-	Vh7183.b7psi
musIGHV192	996469	996775	-	1	V_h7183	5-3b-5, Vh7183.b6, 3:3.9, IGHV5S6*01
RSS	996776	996814				
musIGHV180	1009036	1009340	Pseudo	2	V_h7183	-

V-Segment/ Element	Beginn	Ende	Beson-derheit	VBASE2-Klasse	Familie	Namen
musIGHV190	1010782	1011085	-	1	V_hQ52	2-4b-4, VhQ52.b2
RSS	1011086	1011124				
musIGHV189	1018508	1018812	Relikt	2	-	-
musIGHV181	1024521	1024827	-	1	V_h7183	VHD6.96, IGHV5S2*01
RSS	1024828	1024866				
musIGHV187	1032061	1032365	Relikt	2	-	Vh7183.b3
musIGHV183	1033743	1034046	-	1	V_hQ52	IGHV2S5*01
RSS	1034047	1034085				
musIGHV185	1043505	1043811	-	1	V_h7183	Vh7183.b2
RSS	1043812	1043850				
musIGHV182	1049077	1049383	Pseudo	2	V_h7183	IGHV5S7*01

I.2.B Annotation der D-, J- und konstanten Region

Legende: [1] Hypothetisch. [2] Spacer ist 11 Nukleotide lang. [3] Spacer ist 22 Nukleotide lang. [4] Die Position des Exons wurde durch die cDNAs K0717B09-5N und K0716C06-5N der NIH Maus cDNA Datenbank ermittelt. [5] Keine Konsensus-Spleiß-Sequenz. [6] Kein Konsensus-Spleiß-Donor. [7] Kein Konsensus-PolyA-Signal. [8] Abweichung von der IMGT-Annotation.

Element	Start	Ende	Referenz-Sequenz	Referenz-Literatur	GENSCAN
D-Region					
RSS	1093381	1093408	-	-	-
DFL16.3[1]	1093409	1093431	-	-	-
RSS	1093432	1093459	-	-	-
RSS[2]	1096081	1096107	IMGT:AC073553	Ye 2004	-
DST4.2	1096108	1096124	IMGT:AC073553	Ye 2004	-
RSS	1096125	1096152	IMGT:AC073553	Ye 2004	-
RSS	1145594	1145621	EMBL:J00434	Kurosawa 1982	-
DFL16.1	1145622	1145644	EMBL:J00434	Kurosawa 1982	-
RSS	1145645	1145672	EMBL:J00434	Kurosawa 1982	-
Lücke	1148504	1148603			
RSS	1152455	1152482	EMBL:D13199	Lawler 1987	-
DSP2.9	1152483	1152499	EMBL:D13199	Lawler 1987	-
RSS	1152500	1152527	EMBL:D13199	Lawler 1987	-
RSS	1156774	1156801	EMBL:J00435	Kurosawa 1982	-
DSP2.3	1156802	1156818	EMBL:J00435	Kurosawa 1982	-
RSS	1156818	1156846	EMBL:J00435	Kurosawa 1982	-
RSS[2]	1159318	1159344	-	-	-

Element	Start	Ende	Referenz-Sequenz	Referenz-Literatur	GENSCAN
DST4.3	1159345	1159361	-	-	-
RSS	1159362	1159389	-	-	-
RSS	1160874	1160901	EMBL:J00436	Kurosawa 1982	-
DFL16.2	1160902	1160918	EMBL:J00436	Kurosawa 1982	-
RSS	1160919	1160946	EMBL:J00436	Kurosawa 1982	-
RSS	1165507	1165534	EMBL:J00432	Kurosawa 1982	-
DSP2.5	1165535	1165551	EMBL:J00432	Kurosawa 1982	-
RSS	1165552	1165579	EMBL:J00432	Kurosawa 1982	-
RSS	1170155	1170182	EMBL:J00431	Kurosawa 1982	-
DSP2.2a	1170183	1170199	EMBL:J00431	Kurosawa 1982	-
RSS	1170200	1170227	EMBL:J00431	Kurosawa 1982	-
RSS	1174833	1174860	-	Chang 1992	-
DSP2.11	1174861	1174877	-	Chang 1992	-
RSS	1174878	1174905	-	Chang 1992	-
RSS	1179825	1179852	EMBL:J00431	Kurosawa 1982	-
DSP2.2b	1179853	1179869	EMBL:J00431	Kurosawa 1982	-
RSS	1179870	1179897	EMBL:J00431	Kurosawa 1982	-
RSS	1184485	1184512	EMBL:J00439	Kurosawa 1982	-
DSP2.8	1184513	1184529	EMBL:J00439	Kurosawa 1982	-
RSS	1184530	1184557	EMBL:J00439	Kurosawa 1982	-
RSS	1189617	1189644	EMBL:J00438	Kurosawa 1982	-
DSP2.7	1189645	1189661	EMBL:J00438	Kurosawa 1982	-
RSS	1189662	1189689	EMBL:J00438	Kurosawa 1982	-
RSS[2]	1192293	1192319	EMBL:M23243	Riblet 1993	-
DST4	1192320	1192335	EMBL:M23243	Riblet 1993	-
RSS	1192336	1192363	EMBL:M23243	Riblet 1993	-
RSS	1206820	1206847	EMBL:J00440, EMBL:L32868	Early 1980, Dirkes 1994	-
DQ52	1206848	1206858	EMBL:J00440, EMBL:L32868	Early 1980, Dirkes 1994	-
RSS	1206859	1206886	EMBL:J00440, EMBL:L32868	Early 1980, Dirkes 1994	-
J-Region					
RSS	1207514	1207552	EMBL:V00770	Sakano 1980, Newell 1980	-
JH1	1207553	1207605	EMBL:V00770	Sakano 1980, Newell 1980	-
RSS	1207832	1207870	EMBL:V00770	Sakano 1980, Newell 1980	-
JH2	1207871	1207918	EMBL:V00770	Sakano 1980, Newell 1980	-
RSS	1208215	1208253	EMBL:V00770	Sakano 1980, Newell 1980	-
JH3	1208254	1208301	EMBL:V00770	Sakano 1980, Newell 1980	-

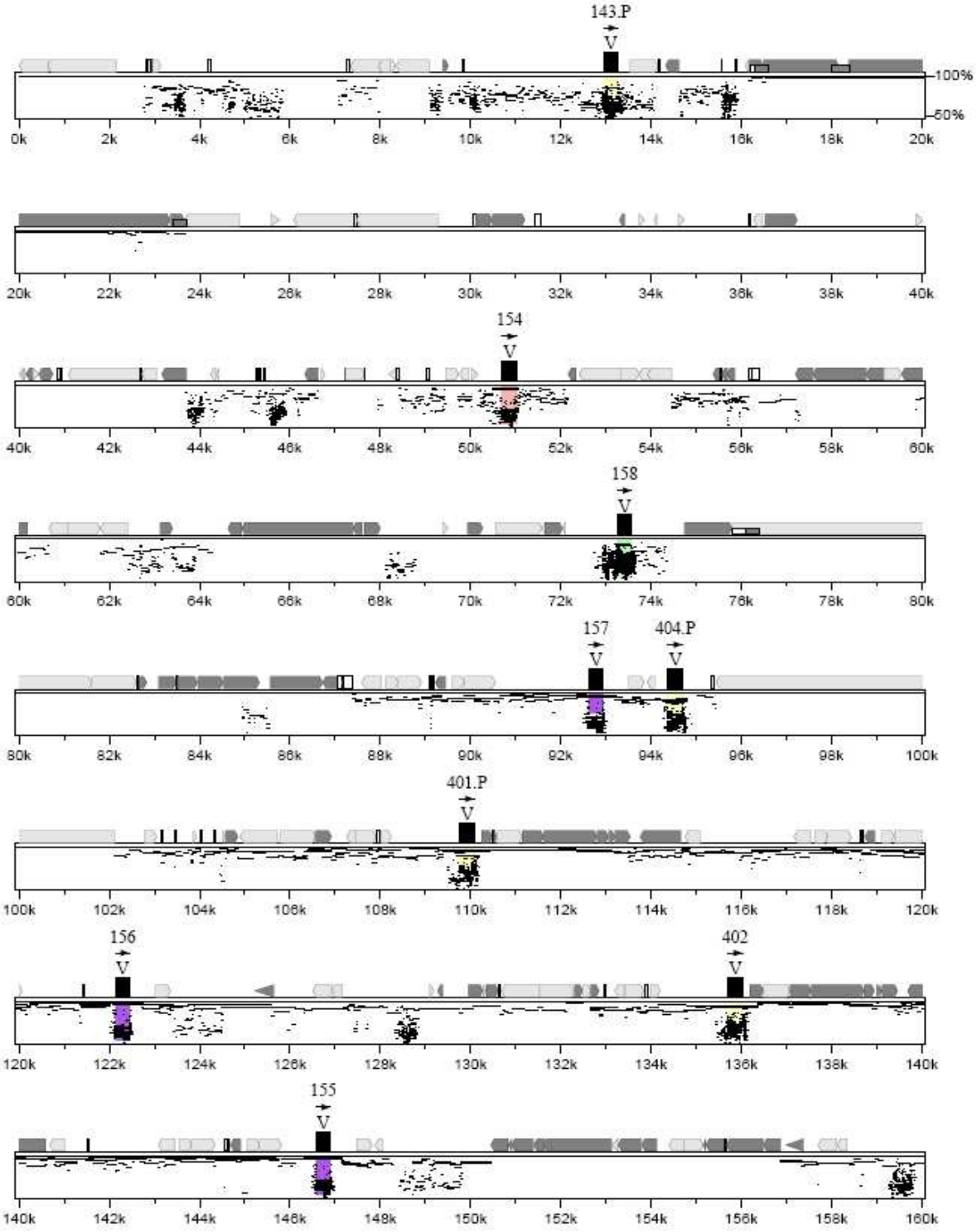
Element	Start	Ende	Referenz-Sequenz	Referenz-Literatur	GENSCAN
RSS[3]	1208782	1208819	EMBL:V00770	Sakano 1980	-
JH4	1208820	1208873	EMBL:V00770	Sakano 1980	-
C-Mu-Region					
Ig-Enhancer-Region[1]	1209791	1210103	EMBL:J00440	Gillies 1983, Banerji 1983	-
I-Exon[4]	1210180	1210633	EMBL:V00771	Sakano 1980	-
Lücke	1212930	1213029			
CH1-Exon	1215719	1216033	EMBL:V00818	Kawakami 1980	+
CH2-Exon	1216144	1216482	EMBL:V00818	Kawakami 1980	+
CH3-Exon	1216762	1217079	EMBL:V00818	Kawakami 1980	+
CH4-M-Exon	1217187	1217519	EMBL:V00818	Kawakami 1980	+|-
CH4-S-Exon	1217187	1217578	EMBL:V00818	Kawakami 1980	+|-
PolyA-Signal	1217681	1217686	EMBL:V00818	Kawakami 1980	-
M1-Exon	1219384	1219499	EMBL:J00444	Rogers 1980	+
M2-Exon	1219518	1219623	EMBL:J00444	Rogers 1980	-
PolyA-Signal[1]	1219869	1219874	EMBL:K02138	Early 1980	-
C-Delta-Region					
CH1-Exon	1222168	1222449	EMBL:V00786	Tucker 1980	-|+
Hinge-Exon	1222826	1222930	EMBL:V00787	Tucker 1980	+
CH3-Exon	1223934	1224254	EMBL:V00788	Tucker 1980	+
CH-S-Exon	1228794	1228858	EMBL:J00450	Tucker 1980	-
PolyA-Signal	1229023	1229028	EMBL:J00450	Tucker 1980	-
CH-M1-Exon	1230516	1230673	EMBL:J00450	Tucker 1980	+
CH-M2-Exon	1230894	1230899	EMBL:J00450	Tucker 1980	-
PolyA-Signal	1231214	1231219	EMBL:J00450	Tucker 1980	-
C-Gamma-3-Region					
I-Exon[5]	1272138	1272621	EMBL:D78343	Akahori 1997	-
CH1-Exon	1277544	1277834	EMBL:D78343	Akahori 1997	+
Hinge-Exon	1278200	1278247	EMBL:D78343	Akahori 1997	+
CH2-Exon	1278347	1278676	EMBL:D78343	Akahori 1997	+
CH3-M-Exon	1278789	1279103	EMBL:D78343	Akahori 1997	+|-
CH3-S-Exon	1278789	1279108	EMBL:D78343	Akahori 1997	+|-
PolyA-Signal	1279180	1279185	EMBL:D78343	Akahori 1997	-
CH-M1-Exon	1280556	1280686	EMBL:D78343	Akahori 1997	+
CH-M2-Exon	1281234	1281314	EMBL:D78343	Akahori 1997	+
PolyA-Signal	1282508	1282513	EMBL:D78343	Akahori 1997	+
C-Gamma-1-Region					
I-Exon[6]	1299698	1300212	EMBL:D78344	Akahori 1997	-

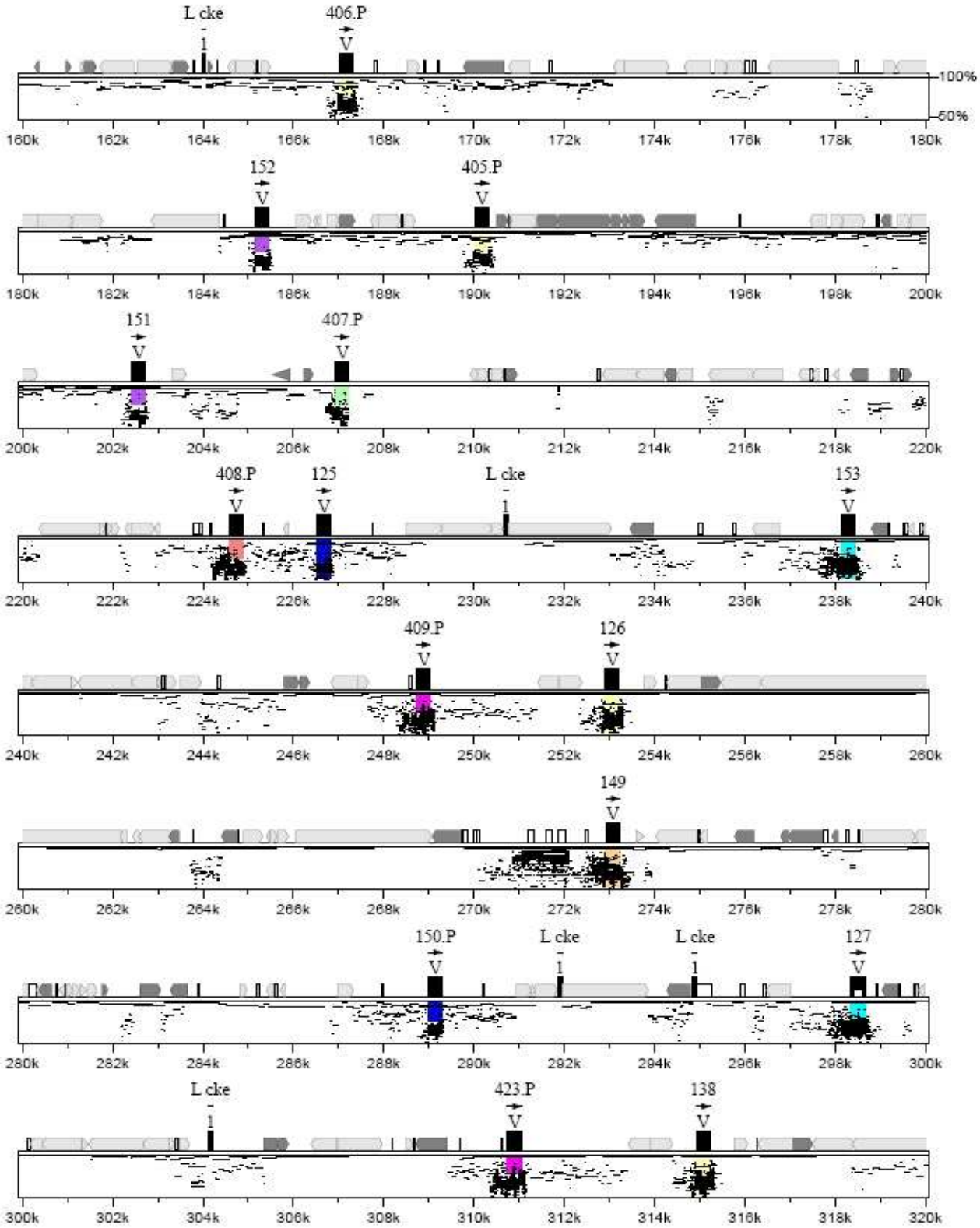
Element	Start	Ende	Referenz-Sequenz	Referenz-Literatur	GENSCAN
CH1-Exon	1310574	1310864	EMBL:D78344	Akahori 1997	+
Hinge-Exon	1311217	1311255	EMBL:D78344	Akahori 1997	-
CH2-Exon	1311359	1311679	EMBL:D78344	Akahori 1997	+
CH3-M-Exon	1311801	1312115	EMBL:J00453	Honjo 1979, Akahori 1997	+
CH3-S-Exon	1311801	1312120	EMBL:J00453	Honjo 1979	+
PolyA-Signal	1312191	1312196	EMBL:J00453	Honjo 1979	-
CH-M1-Exon	1313537	1313667	EMBL:J00454	Yamawaki-Kataoka 1982	+
CH-M2-Exon	1314464	1314544	EMBL:J00455	Yamawaki-Kataoka 1982	-
PolyA-Signal[1]	1315835	1315840	EMBL:D78344	Akahori 1997	-
C gamma 2b-Region					
I-Exon[5]	1326190	1326589	EMBL:D78344	Akahori 1997	-
I-Exon	1326186	1326619	EMBL:L08600	Collins 1993	-
CH1-Exon	1331804	1332094	EMBL:D78344	Akahori 1997	+
Hinge-Exon	1332411	1332476	EMBL:D78344	Akahori 1997	+
CH2-Exon	1332584	1332913	EMBL:D78344	Akahori 1997	+
CH3-M-Exon	1333026	1333340	EMBL:V00763	Akahori 1997	+|-
CH3-S-Exon	1333026	1333345	EMBL:V00763	Akahori 1997	+|-
PolyA-Signal	1333415	1333420	EMBL:D78344	Akahori 1997	
CH-M1-Exon	1334695	1334825	EMBL:J00462, EMBL:D78344	Rogers 1981, Akahori 1997	+
CH-M2-Exon [8]	1335337	1335417	EMBL:J00462, EMBL:D78344	Rogers 1981, Akahori 1997	+
PolyA-Signal[1]	1336238	1336243	-	-	+
Haupt-PolyA-Signal	1336745	1336750	EMBL:M80654	Ward 1992	-
Neben-PolyA-Signal	1336825	1336830	EMBL:M80654	Ward 1992	-
C-Gamma-2a-Region					
I-Exon	1341544	1342294	EMBL:D78344	Akahori 1997	
I-Exon	1341503	1341938	EMBL:L08600	Collins 1993	
CH1-Exon	1348270	1348560	EMBL:V00798, IMGT:X16997	Sikorav 1980, Morgado 1989	+
Hinge-Exon	1348871	1348918	EMBL:V00798, IMGT:X16997	Sikorav 1980, Morgado 1989	-
CH2-Exon	1349028	1349357	EMBL:V00798, IMGT:X16997	Sikorav 1980, Morgado 1989	+
CH3-M-Exon	1349470	1349784	EMBL:V00798, IMGT:X16997	Sikorav 1980, Morgado 1989	+|-
CH3-S-Exon	1349470	1349789	EMBL:V00798, IMGT:X16997	Sikorav 1980, Morgado 1989	+|-
PolyA-Signal[1]	1351302	1351307	EMBL:D78344	Akahori 1997	-
CH-M1-Exon	1351152	1351282	EMBL:J00471	Yamawaki-Kataoka 1982	+

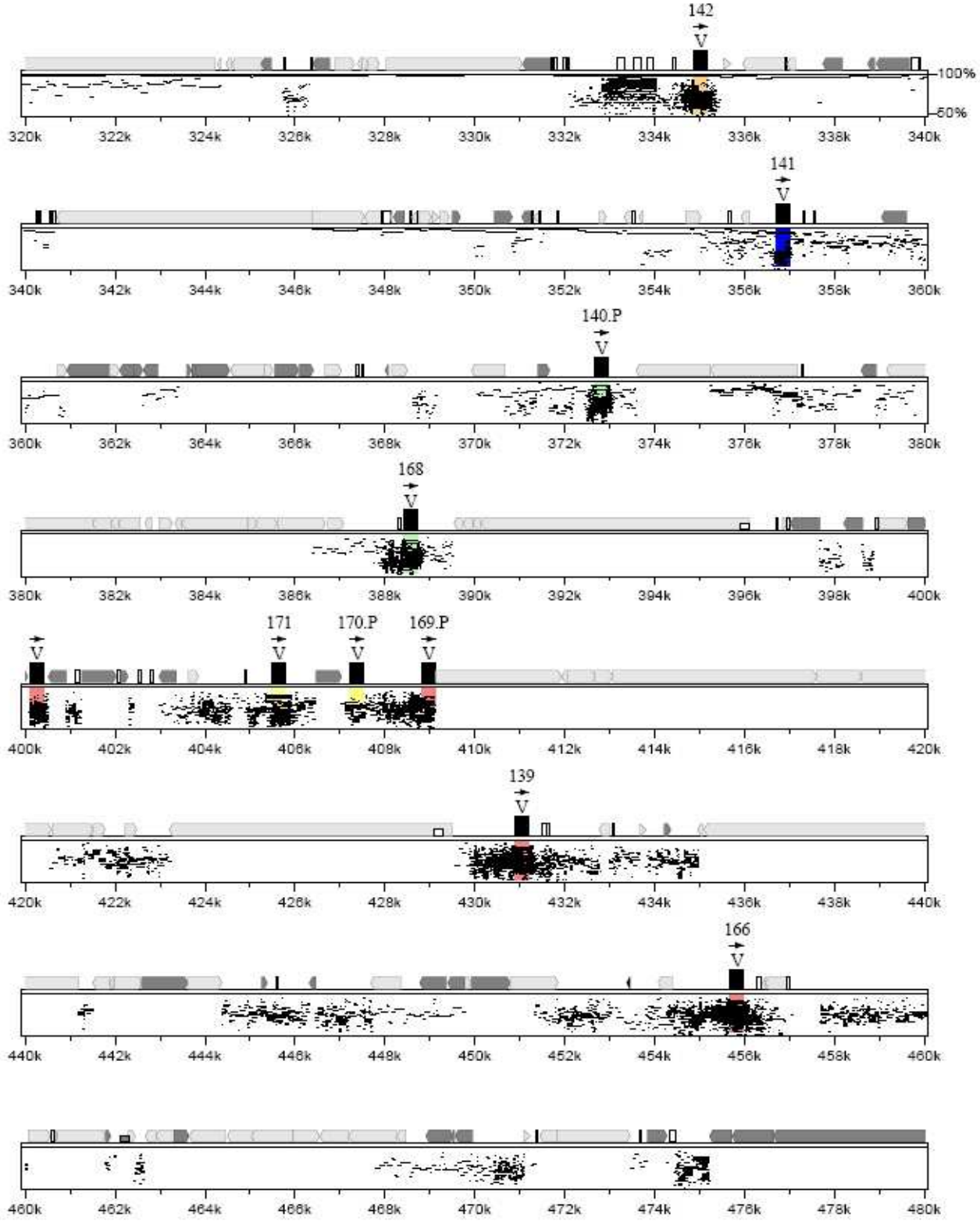
Element	Start	Ende	Referenz-Sequenz	Referenz-Literatur	GENSCAN
CH-M2-Exon[8]	1351791	1351871	EMBL:J00471, EMBL:M35032	Yamawaki-Kataoka 1982, Hall 1989	+
Lücke	1353069	1353168			
Haupt-PolyA-Signal	1353500	1353505	EMBL:M35032	Hall 1989	-
Neben-PolyA-Signal	1353580	1353585	EMBL:M35032	Hall 1989	-
C-Epsilon-Region					
I-Exon	1359443	1359788	EMBL:M31133	Gerondakis 1990	
CH1-Exon	1363970	1364242	EMBL:X01857	Ishida 1982	+
CH2-Exon	1364784	1365104	EMBL:X01857	Ishida 1982	+
CH3-Exon	1365187	1365507	EMBL:X01857	Ishida 1982	+
CH4-M-Exon	1365592	1365918	EMBL:X01857	Ishida 1982	+
CH4-S-Exon	1365592	1365941	EMBL:X01857	Ishida 1982	+|-
PolyA-Signal	1366019	1366024	EMBL:K00689	Ishida 1982	-
CH-M1-Exon	1367618	1367751	EMBL:X03624	Ishida 1982	+
CH-M2-Exon	1367834	1367914	EMBL:X03624	Ishida 1982	-
PolyA-Signal[7]	1368578	1368583	EMBL:X99259	Anand 1997	-
alternatives PolyA-Signal[7]	1369064	1369069	EMBL:X99259	Anand 1997	-
C-Alpha-Region					
I-Exon[6]	1371967	1372423	EMBL:M29011	Radcliffe1990	-
alternatives I-Exon[6]	1372030	1372423	EMBL:M29011	Radcliffe1990	-
CH1-Exon	1377467	1377769	EMBL:J00475, EMBL:D11468	Tucker 1981, Arakawa 1993	+
CH2-Exon	1377999	1378334	EMBL:J00475, EMBL:D11468	Tucker 1981, Arakawa 1993	+
CH3-M-Exon	1378541	1378872	EMBL:J00475, EMBL:D11468	Tucker 1981, Arakawa 1993	+
CH3-S-Exon	1378541	1378932	EMBL:J00475, EMBL:D11468	Tucker 1981, Arakawa 1993	-|+
PolyA-Signal	1378962	1378967	EMBL:J00475	Tucker 1981	+
CH-M-Exon	1381294	1381487	EMBL:K00691	Word 1983	-
PolyA-Signal	1381829	1381834	EMBL:K00691	Word 1983	-

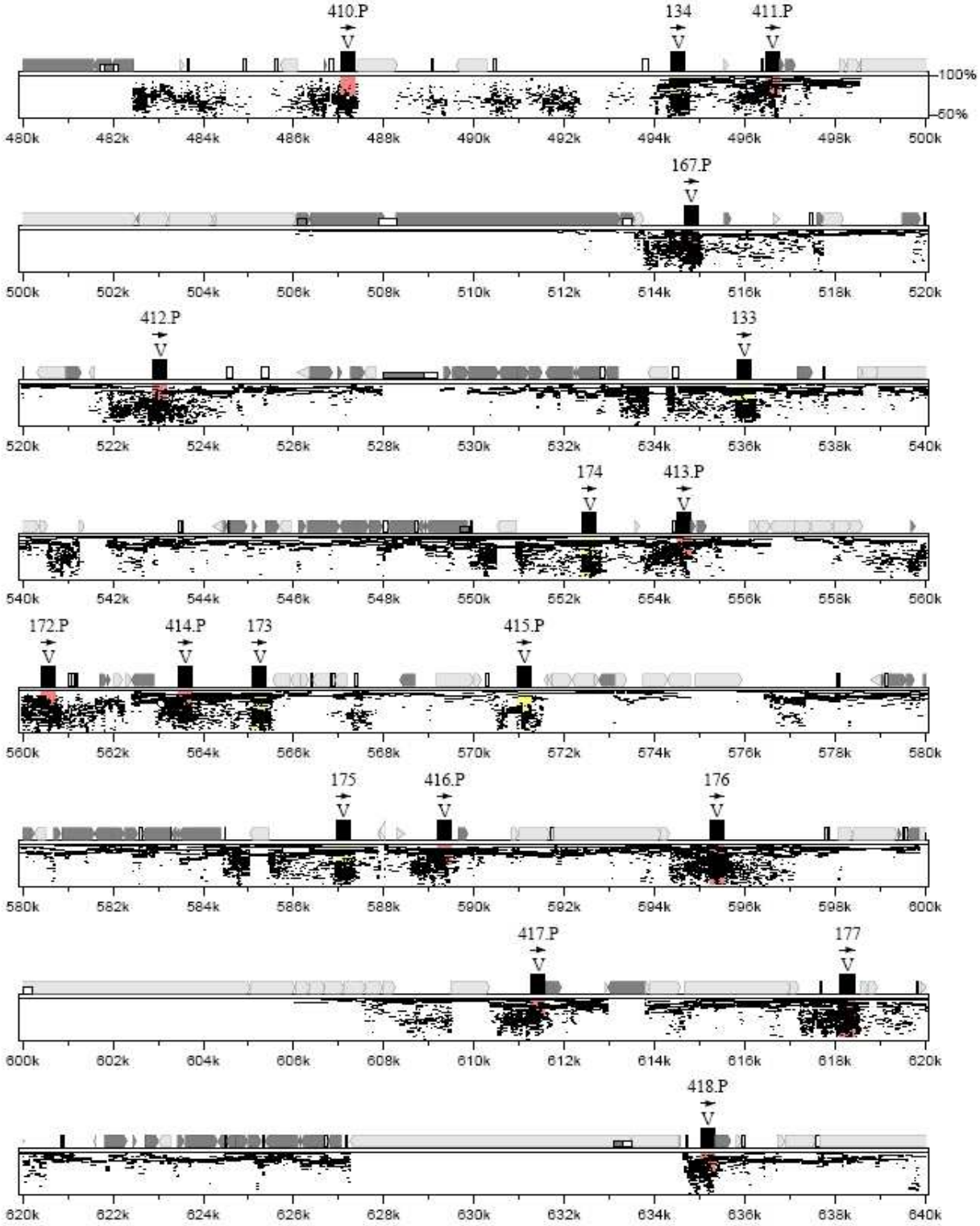
I.3 Ergebnis der PipMaker-Analyse zur Detektion interner homologer Bereiche

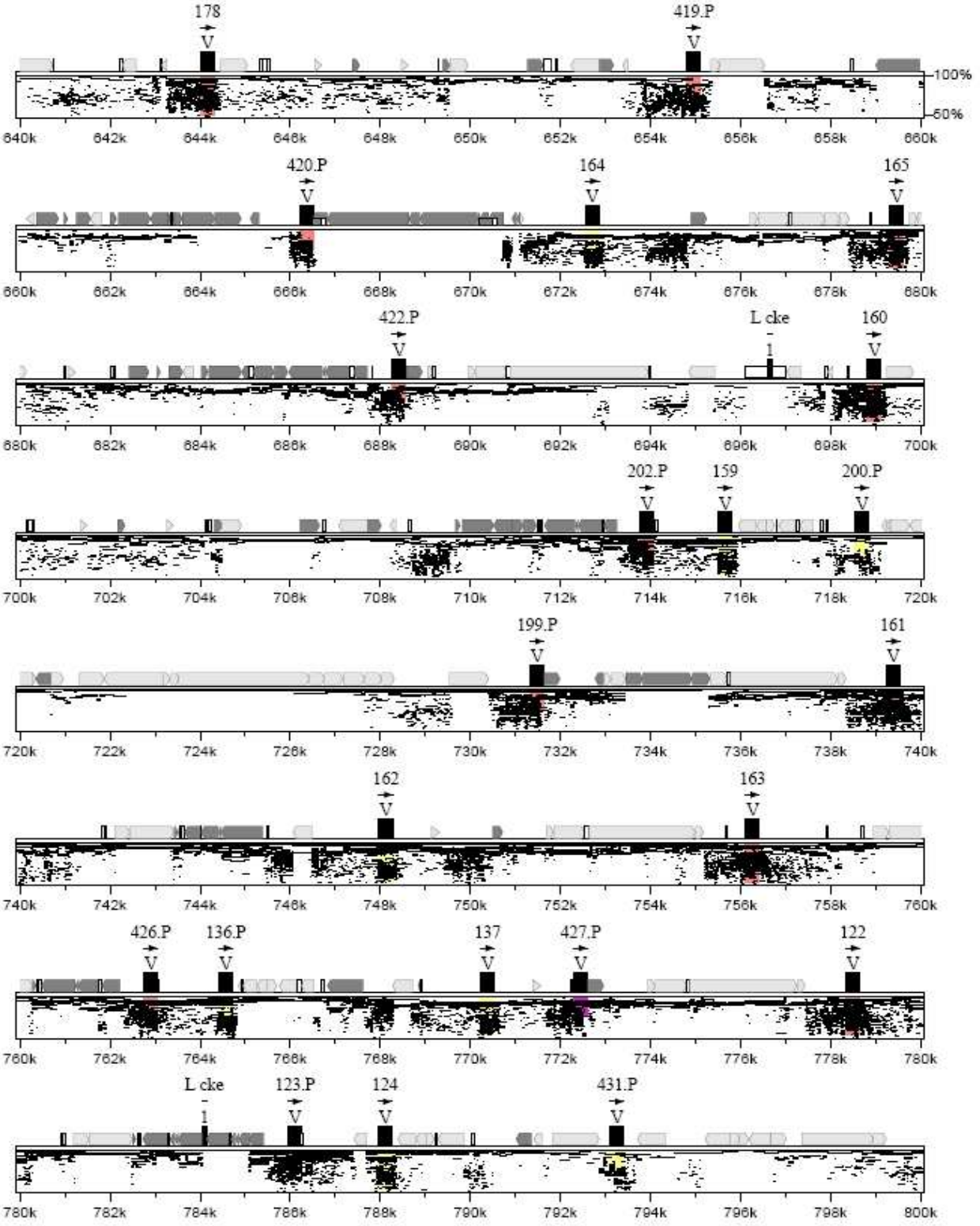
Die Legende der markierten Elemente befindet sich am Ende des Pips.

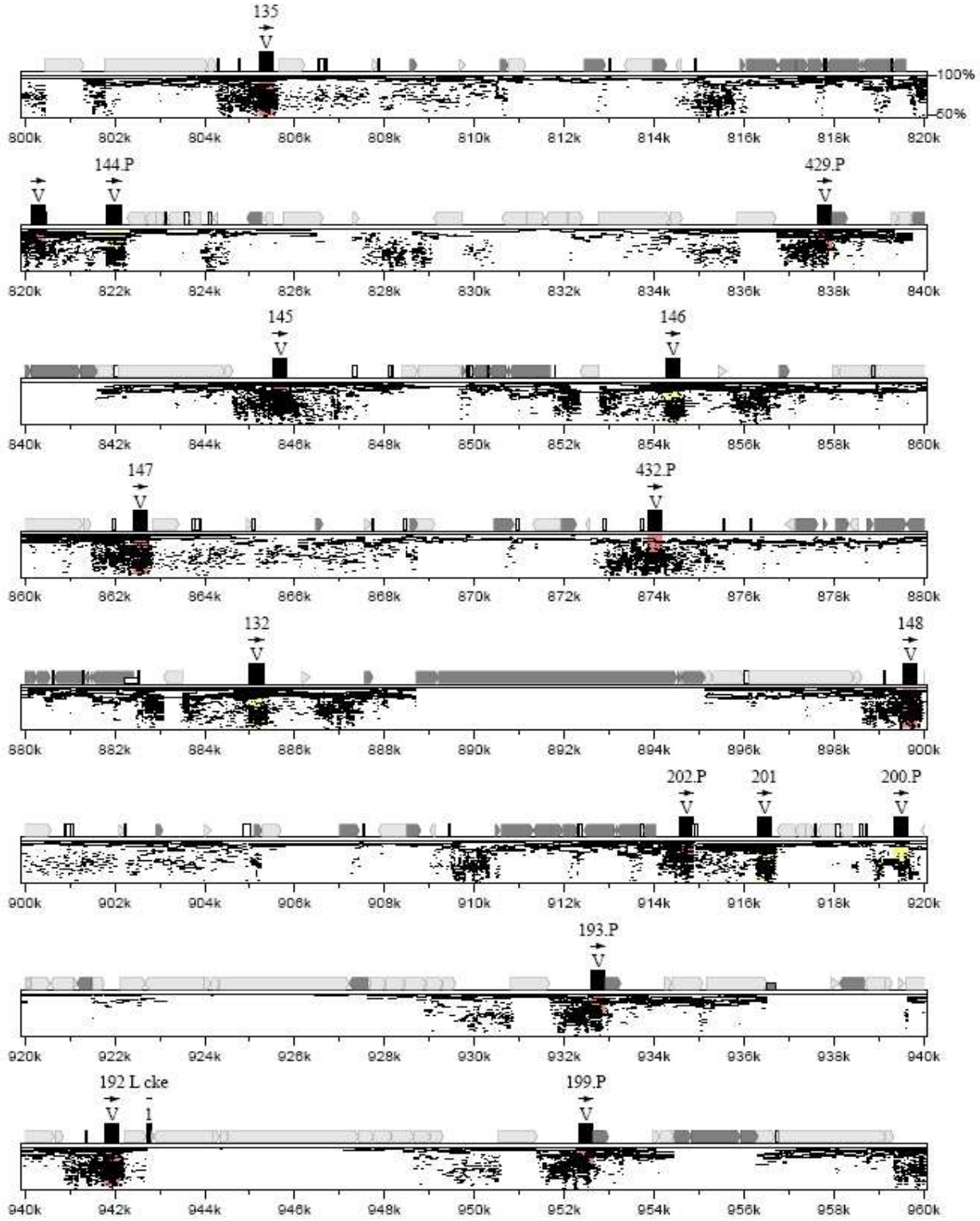


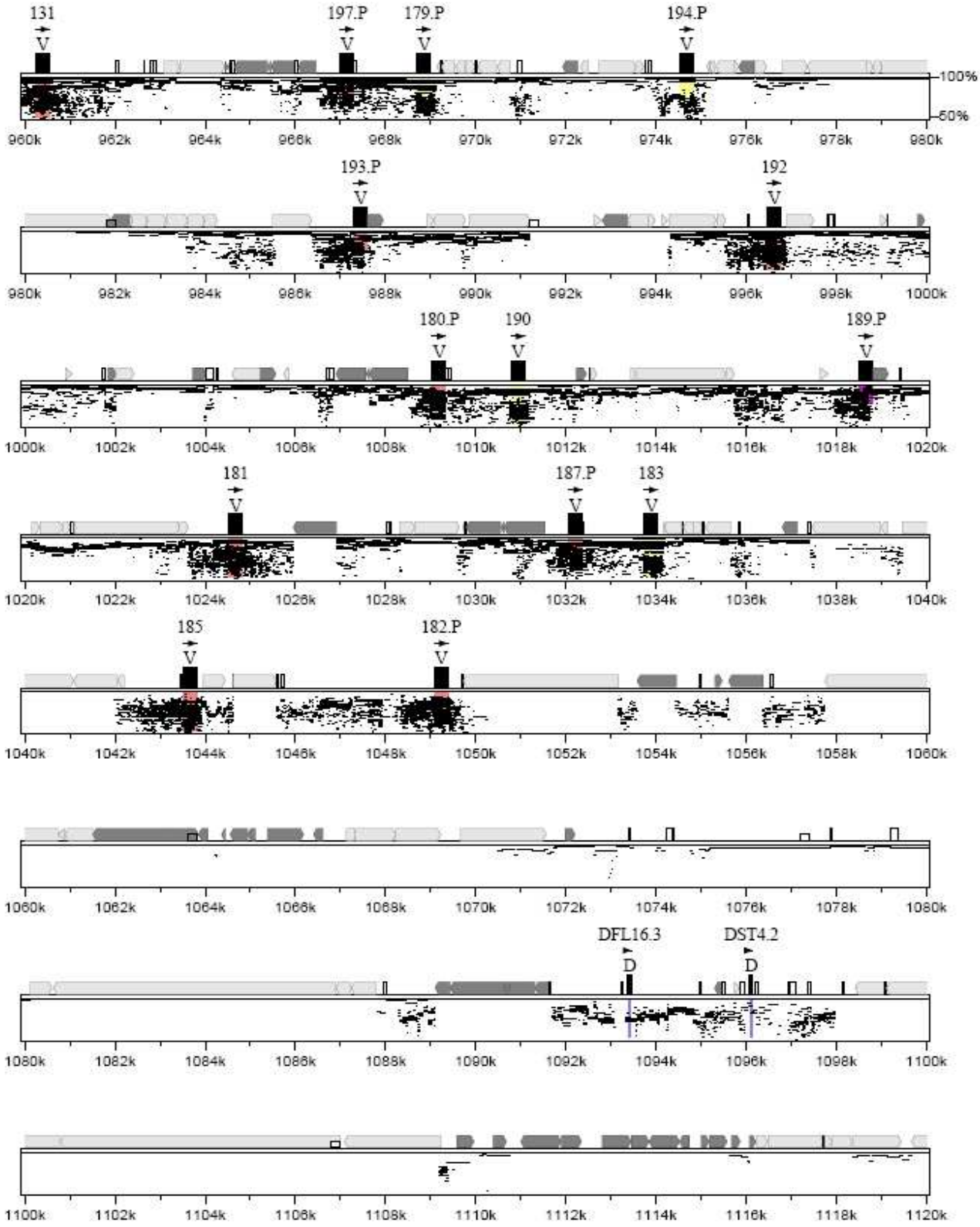

L cke
1
406.P
V
100%
50%
160k 162k 164k 166k 168k 170k 172k 174k 176k 178k 180k
152
V
405.P
V
180k 182k 184k 186k 188k 190k 192k 194k 196k 198k 200k
151
V
407.P
V
200k 202k 204k 206k 208k 210k 212k 214k 216k 218k 220k
408.P
V
125
V
L cke
1
153
V
220k 222k 224k 226k 228k 230k 232k 234k 236k 238k 240k
409.P
V
126
V
240k 242k 244k 246k 248k 250k 252k 254k 256k 258k 260k
149
V
260k 262k 264k 266k 268k 270k 272k 274k 276k 278k 280k
150.P
V
L cke
1
L cke
1
127
V
280k 282k 284k 286k 288k 290k 292k 294k 296k 298k 300k
L cke
1
423.P
V
138
V
300k 302k 304k 306k 308k 310k 312k 314k 316k 318k 320k

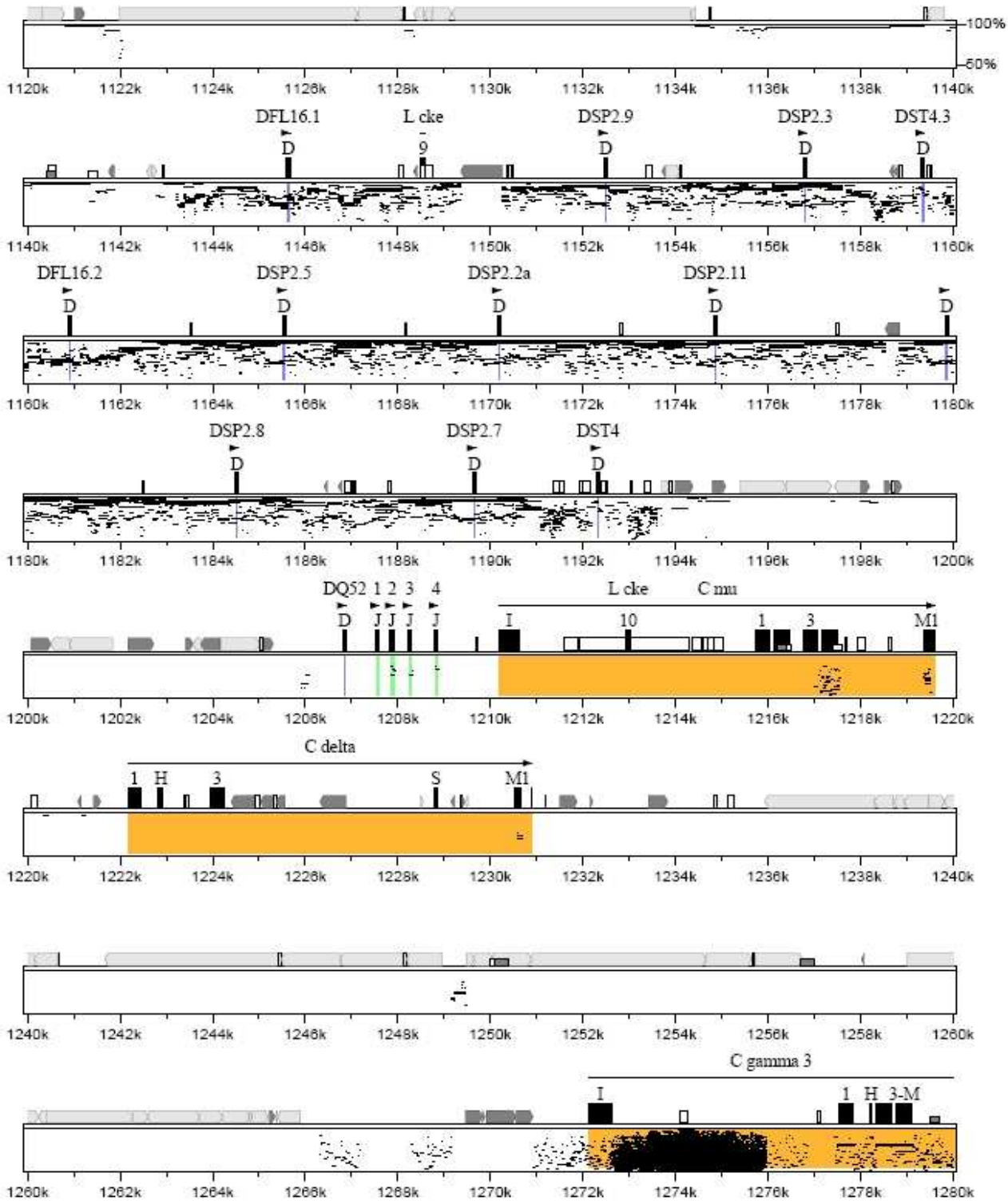

142
100%
60%
320k 322k 324k 326k 328k 330k 332k 334k 336k 338k 340k
141
340k 342k 344k 346k 348k 350k 352k 354k 356k 358k 360k
140.P
360k 362k 364k 366k 368k 370k 372k 374k 376k 378k 380k
168
380k 382k 384k 386k 388k 390k 392k 394k 396k 398k 400k
171
170.P
169.P
400k 402k 404k 406k 408k 410k 412k 414k 416k 418k 420k
139
420k 422k 424k 426k 428k 430k 432k 434k 436k 438k 440k
166
440k 442k 444k 446k 448k 450k 452k 454k 456k 458k 460k
460k 462k 464k 466k 468k 470k 472k 474k 476k 478k 480k

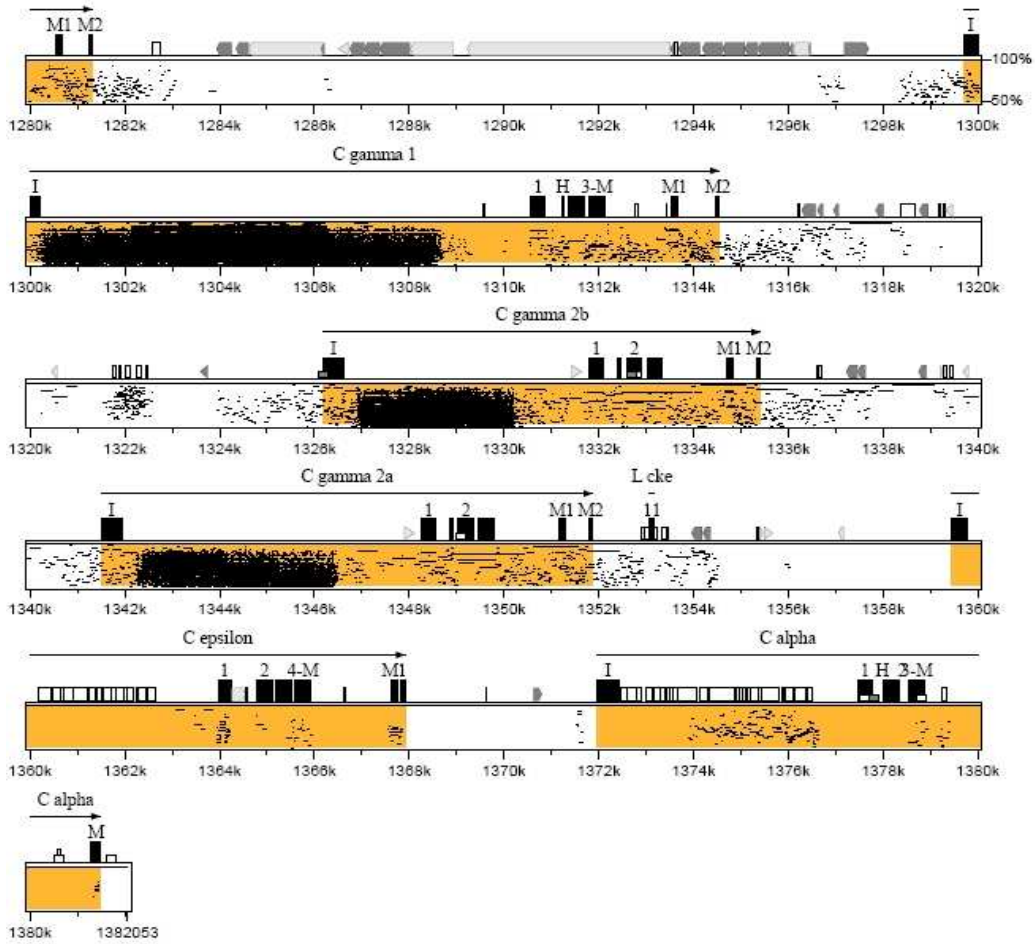

410.P
134
411.P
100%
60%
480k 482k 484k 486k 488k 490k 492k 494k 496k 498k 500k
167.P
500k 502k 504k 506k 508k 510k 512k 514k 516k 518k 520k
412.P
133
520k 522k 524k 526k 528k 530k 532k 534k 536k 538k 540k
174
413.P
540k 542k 544k 546k 548k 550k 552k 554k 556k 558k 560k
172.P
414.P
173
415.P
560k 562k 564k 566k 568k 570k 572k 574k 576k 578k 580k
175
416.P
176
580k 582k 584k 586k 588k 590k 592k 594k 596k 598k 600k
417.P
177
600k 602k 604k 606k 608k 610k 612k 614k 616k 618k 620k
418.P
620k 622k 624k 626k 628k 630k 632k 634k 636k 638k 640k


178
419.P
100%
60%
640k 642k 644k 646k 648k 650k 652k 654k 656k 658k 660k
420.P
164
165
660k 662k 664k 666k 668k 670k 672k 674k 676k 678k 680k
422.P
L cke 1
160
680k 682k 684k 686k 688k 690k 692k 694k 696k 698k 700k
202.P
159
200.P
700k 702k 704k 706k 708k 710k 712k 714k 716k 718k 720k
199.P
161
720k 722k 724k 726k 728k 730k 732k 734k 736k 738k 740k
162
163
740k 742k 744k 746k 748k 750k 752k 754k 756k 758k 760k
426.P
136.P
137
427.P
122
760k 762k 764k 766k 768k 770k 772k 774k 776k 778k 780k
L cke 1
123.P
124
431.P
780k 782k 784k 786k 788k 790k 792k 794k 796k 798k 800k


135
100%
60%
800k 802k 804k 806k 808k 810k 812k 814k 816k 818k 820k
144.P
429.P
820k 822k 824k 826k 828k 830k 832k 834k 836k 838k 840k
145
146
840k 842k 844k 846k 848k 850k 852k 854k 856k 858k 860k
147
432.P
860k 862k 864k 866k 868k 870k 872k 874k 876k 878k 880k
132
148
880k 882k 884k 886k 888k 890k 892k 894k 896k 898k 900k
202.P
201
200.P
900k 902k 904k 906k 908k 910k 912k 914k 916k 918k 920k
193.P
920k 922k 924k 926k 928k 930k 932k 934k 936k 938k 940k
192 L cke
1
199.P
940k 942k 944k 946k 948k 950k 952k 954k 956k 958k 960k


131
197.P
179.P
194.P
100%
50%
960k 962k 964k 966k 968k 970k 972k 974k 976k 978k 980k
193.P
192
980k 982k 984k 986k 988k 990k 992k 994k 996k 998k 1000k
180.P
190
189.P
1000k 1002k 1004k 1006k 1008k 1010k 1012k 1014k 1016k 1018k 1020k
181
187.P
183
1020k 1022k 1024k 1026k 1028k 1030k 1032k 1034k 1036k 1038k 1040k
185
182.P
1040k 1042k 1044k 1046k 1048k 1050k 1052k 1054k 1056k 1058k 1060k
1060k 1062k 1064k 1066k 1068k 1070k 1072k 1074k 1076k 1078k 1080k
DFL16.3
DST4.2
1080k 1082k 1084k 1086k 1088k 1090k 1092k 1094k 1096k 1098k 1100k
1100k 1102k 1104k 1106k 1108k 1110k 1112k 1114k 1116k 1118k 1120k


100%
60%
1120k 1122k 1124k 1126k 1128k 1130k 1132k 1134k 1136k 1138k 1140k
DFL16.1
L cke
9
DSP2.9
DSP2.3
DST4.3
D
1140k 1142k 1144k 1146k 1148k 1150k 1152k 1154k 1156k 1158k 1160k
DFL16.2
DSP2.5
DSP2.2a
DSP2.11
1160k 1162k 1164k 1166k 1168k 1170k 1172k 1174k 1176k 1178k 1180k
DSP2.8
DSP2.7
DST4
1180k 1182k 1184k 1186k 1188k 1190k 1192k 1194k 1196k 1198k 1200k
DQ52 1 2 3 4
J
L cke
C mu
I
10
1
3
M1
1200k 1202k 1204k 1206k 1208k 1210k 1212k 1214k 1216k 1218k 1220k
C delta
H
S
1220k 1222k 1224k 1226k 1228k 1230k 1232k 1234k 1236k 1238k 1240k
1240k 1242k 1244k 1246k 1248k 1250k 1252k 1254k 1256k 1258k 1260k
C gamma 3
3-M
1260k 1262k 1264k 1266k 1268k 1270k 1272k 1274k 1276k 1278k 1280k



Legende:

A Repetitive Sequenzen der Repeatmasker-Analyse

- Gene
- Exon
- Simple
- MIR
- Other SINE
- LINE1
- LINE2
- LTR
- Other repeat
- CpG/GpC≥0.60
- CpG/GpC≥0.75

B Farbige Markierung der kodierenden Sequenzen und Bereiche der konstanten Regionen

- VHJ606 : Pink
- VH15 : LightRed
- VHGam3 : Purple
- VH11 : Cyan
- VH3660 : LightYellow
- VHX24 : LightOrange
- VHSm7 : DarkBlue
- VHS107 : LightGreen
- VHQ52 : Yellow
- VH7183 : Red
- D_Segmente : Blue
- J_Segmente : Green
- C_Isotypen : Orange
- L : Black

II Verzeichnisse

II.1 Verzeichnis der Abbildungen

II.2 Verzeichnis der Tabellen

II.3 Verzeichnis der Abkürzungen

Kurzform	Bedeutung
AC	Accession Number
AID	Activation-Induced cytidine Deaminase
BAC	Bacterial Artificial Chromosome
BLAST	Basic Local Alignment Search Tool
bp	Basenpaare (Nukleotide)
CDR	Complementary Determining Region
cDNA	complementary DNA
CDS	CoDing Sequence
DAS	Distributed Annotation Server
DNA	DesoxyRibonucleic Acid
EBI	European Bioinformatic Institute
EMBL	European Molecular Biology Laboratory
EMVEC	EMBL VECtor database; auch EVEC
FR	Framework Region
FTP	File Transfer Protocol
HMM	Hidden Markov Model
GB	Giga Byte
GBF	Gesellschaft für Biotechnologische Forschung
HSP	High Scoring Pair
HTG	High Throughput Genomic sequences
HVR	Hyper Variable Region
ID	IDentity Number
IgH	ImmunoGlobulin Heavy chain
IgK	ImmunoGlobulin Kappa chain
IgL	ImmunoGlobulin Lambda chain
IgSF	ImmunoGlobulin SuperFamily
IMGT	ImMunoGeneTics, The International Immunogenetics Information System
kb	KiloBasen (Längenangabe einer Nukleotidsequenz)
LIGM	Laboratoire d'ImmunoGénétique Moléculaire
LINE	Long Interspersed Nucleotide Element
LTR	Long Terminal Repeat
MAR	Matrix Attachment Region
Mb	MegaBasen (Längenangabe einer Nukleotidsequenz)

Kurzform	Bedeutung
MB	MegaBytes
MGI	Mouse Genome Informatics
MGNC	Mouse Genomic Nomenclature Committee
NCBI	National Center for Biotechnology Information
NHEJ	Non Homologous End Joining
PHP	PHP: Hypertext Preprocessor
Pip	Percent Identity Plot
RDBMS	Relational Database Management System
RNA	RiboNucleic Acid
RSS	Recombination Signal Sequence
SINE	Short Interspersed Nucleotide Element
SRS	Sequence Retrieval System
TFBS	Transkriptionsfaktor-Bindungsstelle
TPIMS	Torrey Pines Institute for Molecular Studies
TrEMBL	TRanslated EMBL
UniProtKB	UniProt Kowledgebase
WGS	Whole Genome Shotgun Sequences

II.4 Verzeichnis der Webdienste und Datenbanken

Name	Adresse	Funktion
ABG	http://www.ibt.unam.mx/vir/V_mice.html	Almagro Keimbahn-V-Gen-Sammlung
DNAPLOT	http://www.dnaplot.de/ http://www.dnaplot.org/	Alignment, Analyse und Formatierung von V-Gen- und anderen DNA-Sequenzen
EBI	http://www.ebi.ac.uk/	Europäisches Bioinformatik-Institut: zahlreiche Datenbanken und Tools
EMBL-Bank	http://www.ebi.ac.uk/embl/ ftp://ftp.ebi.ac.uk/pub/databases/embl/	Europas primäre Nukleotid-Sequenz-Datenbank
EMVEC	http://www.ebi.ac.uk/blastall/vectors.html	BLAST-Suche zur Identifizierung von Vektor-Sequenz
GAP4	http://staden.sourceforge.net/	Assemblierung genomischer Sequenzen
HGNC	http://www.gene.ucl.ac.uk/nomenclature	Nomenklatur-Kommittee des Human-Genom-Projektes (HUGO)
IMGT	http://imgt.cines.fr/	Internationales Immunogenetik-Informationssystem
Mouse Genome Informatics	http://www.informatics.jax.org/	Maus-Genom-Seite des Jackson Labors
Kabat-DB am EBI	ftp://ftp.ebi.ac.uk/pub/databases/kabat/Rel5.0/	letzte freie Version der Kabat-Datenbank
MGNC	http://www.informatics.jax.org/mgihome/nomen/	Nomenklatur-Kommittee des Maus-Genom-Projektes
NCBI	http://www.ncbi.nlm.nih.gov/	Nationales Informationszentrum für Biotechnologie, USA
PipMaker	http://bio.cse.psu.edu/pipmaker	Programme zur Berechnung ähnlicher Bereiche in 2 oder mehr Input-Sequenzen.
SRS	http://srs.ebi.ac.uk/	Zugang zu Datenbanken des EBI und anderen Datenbanken
Repeatmasker	http://www.repeatmasker.org/	Programm zur Maskierung von repetitiver DNA
Vbase	http://vbase.mrc-cpe.cam.ac.uk/	humane Keimbahn-V-Gen-Datenbank
VBASE2	http://www.vbase2.org/	Integrative Keimbahn-V-Gen-Datenbank
VBASE2 DAS	http://www.dnaplot.org/das/human http://www.dnaplot.org/das/mouse	Ensembl DAS-Server der VBASE2-V-Gene

II.5 Literaturverzeichnis

Akahori, Y., Kurosawa, Y. (1997): Nucleotide sequences of all the gamma gene loci of murine immunoglobulin heavy chains. Genomics 41(1): 100-4.

Al-Lazikani, B., Lesk, A.M., Chothia, C. (1997): Standard conformations for the canonical structures of immunoglobulins. J. Mol. Biol. 273(4): 927-48.

Almagro, J.C., Hernandez, I., del Carmen Ramirez M., Vargas-Madrazo, E. (1997): The differences between the structural repertoires of VH germ-line gene segments of mice and humans: implication for the molecular mechanism of the immune response. Mol. Immunol. 34(16-17): 1199-214.

Alt, F.W., Blackwell, T.K., Yancopoulos, G.D. (1987): Development of the primary antibody repertoire. Science 238(4830): 1079-87.

Alt, F.W., Oltz, E.M., Young, F., Gorman, J., Taccioli, G., Chen, J. (1992): VDJ recombination. Immunol. Today. 13(8): 306-14.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. (1997): Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic. Acids. Res. 25(17): 3389-402.

Anand, S., Batista, F.D., Tkach, T., Efremov, D.G., Burrone, O.R. (1997): Multiple transcripts of the murine immunoglobulin epsilon membrane locus are generated by alternative splicing and differential usage of two polyadenylation sites. Mol. Immunol. 34(2): 175-83.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., Yeh, L.S. (2004): UniProt: the Universal Protein knowledgebase. Nucleic. Acids. Res. 32(Database issue): D115-9.

Arakawa, H., Buerstedde, J.M. (2004): Immunoglobulin gene conversion: insights from bursal B cells and the DT40 cell line. Dev. Dyn. 229(3): 458-64.

Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., Yeh, L.S. (2005): The Universal Protein Resource (UniProt). Nucleic. Acids. Res. 33(Database issue): D154-9.

Banerji, J., Olson, L., Schaffner, W. (1983): A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. Cell 33(3): 729-40.

Bassing, C.H., Swat, W., Alt, F.W. (2002): The mechanism and regulation of chromosomal V(D)J recombination. Cell 109 Suppl: S45-55.

Berek, C., Berger, A., Apel, M. (1991): Maturation of the immune response in germinal centers. Cell 67(6): 1121-9.

Bergman, Y., Cedar, H. (2004): A stepwise epigenetic process controls immunoglobulin allelic exclusion. Nat. Rev. Immunol. 4(10): 753-61.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M. (2003): The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic. Acids. Res. 31(1): 365-70.

Bolland, D.J., Wood, A.L., Johnston, C.M., Bunting, S.F., Morgan, G., Chakalova, L., Fraser, P.J., Corcoran, A.E. (2004): Antisense intergenic transcription in V(D)J recombination. Nat. Immunol. 5(6): 630-7.

Bonfield, J.K., Smith, K., Staden, R. (1995): A new DNA sequence assembly program. Nucleic. Acids. Res. 23(24): 4992-9.

Brandt, V.L., Roth, D.B. (2004): V(D)J recombination: how to tame a transposase. Immunol. Rev. 200: 249-60.

Brodeur, P.H., Riblet, R. (1984): The immunoglobulin heavy chain variable region (Igh-V) locus in the mouse. I. One hundred Igh-V genes comprise seven families of homologous genes. Eur. J. Immunol. 14(10): 922-30.

Bruggemann, M., Free, J., Diamond, A., Howard, J., Cobbold, S., Waldmann, H. (1986): Immunoglobulin heavy chain locus of the rat: striking homology to mouse antibody genes. Proc. Natl. Acad. Sci. U. S. A. 83(16): 6075-9.

Burge, C., Karlin, S. (1997): Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268(1): 78-94.

Cannon, J.P., Haire, R.N., Rast, J.P., Litman, G.W. (2004): The phylogenetic origins of the antigen-binding receptors and somatic diversification mechanisms. Immunol. Rev. 200: 12-22.

Chang, Y., Paige, C.J., Wu, G.E. (1992): Enumeration and characterization of DJH structures in mouse fetal liver. EMBO. J. 11(5): 1891-9.

Chaudhuri, J., Alt, F.W. (2004): Class-switch recombination: interplay of transcription, DNA deamination and DNA repair. Nat. Rev. Immunol. 4(7): 541-52.

Chevillard, C., Ozaki, J., Herring, C.D., Riblet, R. (2002): A three-megabase yeast artificial chromosome contig spanning the C57BL mouse Igh locus. J. Immunol. 168(11): 5659-66.

Chimpanzee Sequencing Consortium. (2005): Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437(7055): 69-87.

Chothia, C., Lesk, A.M. (1987): Canonical structures for the hypervariable regions of immunoglobulins. J. Mol. Biol. 196(4): 901-17.

Collins, J.T., Dunnick, W.A. (1993): Germline transcripts of the murine immunoglobulin gamma 2a gene: structure and induction by IFN-gamma. Int. Immunol. 5(8): 885-91.

Cook, G.P., Tomlinson, I.M. (1995): The human immunoglobulin VH repertoire. Immunol. Today. 16(5): 237-42.

Corcoran, A.E. (2005): Immunoglobulin locus silencing and allelic exclusion. Semin. Immunol. 17(2): 141-54.

Cowell, L.G., Davila, M., Kepler, T.B., Kelsoe, G. (2002): Identification and utilization of arbitrary correlations in models of recombination signal sequences. Genome. Biol. 3(12): RESEARCH0072.

Dammers, P.M., Kroese, F.G. (2001): Evolutionary relationship between rat and mouse immunoglobulin IGHV5 subgroup genes (PC7183) and human IGHV3 subgroup genes. Immunogenetics 53(6): 511-7.

Davila, M., Foster, S., Kelsoe, G., Yang, K. (2001): A role for secondary V(D)J recombination in oncogenic chromosomal translocations? Adv. Cancer. Res. 81: 61-92.

Davis, M.M., Kim, S.K., Hood, L.E. (1980): DNA sequences mediating class switching in alpha-immunoglobulins. Science 209(4463): 1360-5.

De Bono, B., Chothia, C. (2003): Exegesis: a procedure to improve gene predictions and its use to find immunoglobulin superfamily proteins in the human and mouse genomes. Nucleic. Acids. Res. 31(21): 6096-103.

De Bono, B., Madera, M., Chothia, C. (2004): VH gene segments in the mouse and human genomes. J. Mol. Biol. 342(1): 131-43.

Dewannieux, M., Esnault, C., Heidmann, T. (2003): LINE-mediated retrotransposition of marked Alu sequences. Nat. Genet. 35(1): 41-8.

Dewannieux, M., Heidmann, T. (2005): L1-mediated retrotransposition of murine B1 and B2 SINEs recapitulated in cultured cells. J. Mol. Biol. 349(2): 241-7.

Dildrop, R. (1986): Classification of mouse VH sequences. In: Handbook of experimental immunology in four volumes. Volume 3: Genetics and molecular immunology.

Diaz, M., Lawrence, C. (2005): An update on the role of translesion synthesis DNA polymerases in Ig hypermutation. Trends. Immunol. 26(4): 215-20.

Dirkes, G., Kohler, G., Kottmann, A.H. (1994): Sequence and structure of the mouse IgH DQ52 5' region. Immunogenetics 40(5): 379.

Early, P., Huang, H., Davis, M., Calame, K., Hood, L. (1980): An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: VH, D and JH. Cell 19(4): 981-92.

Eason, D.D., Cannon, J.P., Haire, R.N., Rast, J.P., Ostrov, D.A., Litman, G.W. (2004): Mechanisms of antigen receptor evolution. Semin. Immunol. 16(4): 215-26.

Ehlich, A., Martin, V., Müller, W., Rajewsky, K. (1994): Analysis of the B-cell progenitor compartment at the level of single cells. Curr. Biol. 4(7): 573-83.

Else, K., Wakelin, D. (1989): Genetic variation in the humoral immune responses of mice to the nematode Trichuris muris. Parasite. Immunol. 11(1): 77-90.

Engel, H., Ruhl, H., Benham, C.J., Bode, J., Weiss, S. (2001): Germ-line transcripts of the immunoglobulin lambda J-C clusters in the mouse: characterization of the initiation sites and regulatory elements. Mol. Immunol. 38(4): 289-302.

Esser, C., Radbruch, A. (1990): Immunoglobulin class switching: molecular and cellular analysis. Annu. Rev. Immunol. 8: 717-35.

Feeney, A.J., Riblet, R. (1993): DST4: a new, and probably the last, functional DH gene in the BALB/c mouse. Immunogenetics 37(3): 217-21.

Feeney, A.J. (2000): Factors that influence formation of B cell repertoire. Immunol. Res. 21(2-3): 195-202.

Feeney, A.J., Tang, A., Ogwaro, K.M. (2000): B-cell repertoire formation: role of the recombination signal sequence in non-random V segment utilization. Immunol. Rev. 175: 59-69.

Feeney, A.J., Goebel, P., Espinoza, C.R. (2004): Many levels of control of V gene rearrangement frequency. Immunol. Rev. 200: 44-56.

Filpula, D., McGuire, J. (1999): Single-chain Fv designs for protein, cell and gene therapeutics. Exp. Opin. Ther. Patents. 9(3): 231-245.
Flajnik, M.F. (2002): Comparative analyses of immunoglobulin genes: surprises and portents. Nat. Rev. Immunol. 2(9): 688-98.

Flajnik, M.F., Du Pasquier. (2004): Evolution of innate and adaptive immunity: can we draw a line? Trends. Immunol. 25(12): 640-4.

Fowell, D.J., Locksley, R.M. (1999): Leishmania major infection of inbred mice: unmasking genetic determinants of infectious diseases. Bioessays 21(6): 510-8.

Franklin, A., Blanden, R.V. (2004): On the molecular mechanism of somatic hypermutation of rearranged immunoglobulin genes. Immunol. Cell. Biol. 82(6): 557-67.

Frazer, J.K., Capra, J.D. (1999): Immunoglobulins: Structure and Function. In: Fundamental Immunology. Paul (ed); Lippincott-Raven, Philadeplphia; 37-74.

Fuxa, M., Skok, J., Souabni, A., Salvagiotto, G., Roldan, E., Busslinger, M. (2004): Pax5 induces V-to-DJ rearrangements and locus contraction of the immunoglobulin heavy-chain gene. Genes. Dev. 18(4): 411-22.

Gerondakis, S. (1990): Structure and expression of murine germ-line immunoglobulin epsilon heavy chain transcripts induced by interleukin 4. Proc. Natl. Acad. Sci. U. S. A. 87(4): 1581-5.

Gerondakis, S., Gaff, C., Goodman, D.J., Grumont, R.J. (1991): Structure and expression of mouse germline immunoglobulin gamma 3 heavy chain transcripts induced by the mitogen lipopolysaccharide. Immunogenetics 34(6): 392-400.

Gillies, S.D., Morrison, S.L., Oi, V.T., Tonegawa, S. (1983): A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. Cell 33(3): 717-28.

Goebel, P., Montalbano, A., Ayers, N., Kompfner, E., Dickinson, L., Webb, C.F., Feeney, A.J. (2002): High frequency of matrix attachment regions and cut-like protein x/CCAAT-displacement protein and B cell regulator of IgH transcription binding sites flanking Ig V region genes. J. Immunol. 169(5): 2477-87.

Goldmit, M., Bergman, Y. (2004): Monoallelic gene expression: a repertoire of recurrent themes. Immunol. Rev. 200: 197-214.

Goldmit, M., Ji, Y., Skok, J., Roldan, E., Jung, S., Cedar, H., Bergman, Y. (2005): Epigenetic ontogeny of the Igk locus during B cell development. Nat. Immunol. 6(2): 198-203.

Green, M.C. (1979): Genetic nomenclature for the immunoglobulin loci of mouse. Immunogenetics 8: 89-97.

Gu, H., Kitamura, D., Rajewsky, K. (1991): B cell development regulated by gene rearrangement: arrest of maturation by membrane-bound D mu protein and selection of DH element reading frames. Cell 65(1): 47-54.

Gu, H., Tarlinton, D., Müller, W., Rajewsky, K., Forster, I. (1991): Most peripheral B cells in mice are ligand selected. J. Exp. Med. 173(6): 1357-71.

Haines, B.B., Angeles, C.V., Parmelee, A.P., McLean, P.A., Brodeur, P.H. (2001): Germline diversity of the expressed BALB/c VhJ558 gene family. Mol. Immunol. 38(1): 9-18.

Hall, B., Milcarek, C. (1989): Sequence and polyadenylation site determination of the murine immunoglobulin gamma 2a membrane 3' untranslated region. Mol. Immunol. 26(9): 819-26.

Han, J.S., Szak, S.T., Boeke, J.D. (2004): Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature 429(6989): 268-74.

Hayakawa, K., Hardy, R.R. (2000): Development and function of B-1 cells. Curr. Opin. Immunol. 12(3): 346-53.

Henderson, A., Calame, K. (1998): Transcriptional regulation during B cell development. Annu. Rev. Immunol. 16: 163-200.

Herring, C.D., Chevillard, C., Johnston, S.L., Wettstein, P.J., Riblet, R. (1998): Vector-hexamer PCR isolation of all insert ends from a YAC contig of the mouse Igh locus. Genome. Res. 8(6): 673-81.

Hesslein, D.G., Schatz, D.G. (2001): Factors and forces controlling V(D)J recombination. Adv. Immunol. 78: 169-232.

Hofker, M.H., Walter, M.A., Cox, D.W. (1989): Complete physical map of the human immunoglobulin heavy chain constant region gene complex. Proc. Natl. Acad. Sci. U. S. A. 86(14): 5567-71.

Honjo, T. (1983): Immunoglobulin genes. Annu. Rev. Immunol. 1: 499-528.

Honjo T, Alt FW. (ed) (1995): Immunoglobulin Genes. Academic Press Limited, London.

Honjo, T., Kataoka, T. (1978): Organization of immunoglobulin heavy chain genes and allelic deletion model. Proc. Natl. Acad. Sci. USA 75(5): 2140-4.

Honjo, T., Muramatsu, M., Fagarasan, S. (2004): AID: how does it aid antibody diversity? Immunity 20(6): 659-68.

Honjo, T., Nagaoka, H., Shinkura, R., Muramatsu, M. (2005): AID to overcome the limitations of genomic information. Nat. Immunol. 6(7): 655-61.

Honjo, T., Obata, M., Yamawaki-Katoaka, Y., Kataoka, T., Kawakami, T., Takahashi, N., Mano, Y. (1979): Cloning and complete nucleotide sequence of mouse immunoglobulin gamma 1 chain gene. Cell 18(2): 559-68.

Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X.M., Gilbert, J., Hammond, M., Herrero, J., Hotz, H., Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Kokocinsci, F., London, D., Longden, I., McVicker, G., Melsopp, C., Meidl, P., Potter, S., Proctor, G., Rae, M., Rios, D., Schuster, M., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C., Birney, E. (2005): Ensembl 2005. Nucleic. Acids. Res. 33(Database issue): D447-53.

Hutchison, C.A.III, Hardies, S.C., Loeb, D.D., Shehee, W.R., Edgell, M.H. (1989). LINEs and related retroposons: long interspersed repeated sequences in the eucaryotic genome. In: Mobile DNA. Berg DE, Howe (ed); MM American Society for Microbiology, Washington, D.C.; 593-617.

Ishida, N., Ueda, S., Hayashida, H., Miyata, T., Honjo, T. (1982): The nucleotide sequence of the mouse immunoglobulin epsilon gene: comparison with the human epsilon gene sequence. EMBO. J. 1(9): 1117-23.

Johnson, G., Wu, T.T. (2001): Kabat Database and its applications: future directions. Nucleic. Acids. Res. 29(1): 205-6.

Johnson, K., Shapiro-Shelef, M., Tunyaplin, C., Calame, K. (2005): Regulatory events in early and late B-cell differentiation. Mol. Immunol. 42(7): 749-61.

Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van, d.e.n., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Gamble, J., Diez, F.G., Harte, N., Kulikova, T., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., McHale, M., Nardone, F., Silventoinen, V., Sobhany, S., Stoehr, P., Tuli, M.A., Tzouvara, K., Vaughan, R., Wu, D., Zhu, W., Apweiler, R. (2005): The EMBL Nucleotide Sequence Database. Nucleic. Acids. Res. 33(Database issue): D29-33.

Kataoka, T., Kawakami, T., Takahashi, N., Honjo, T. (1980): Rearrangement of immunoglobulin gamma 1-chain gene and mechanism for heavy-chain class switch. Proc. Natl. Acad. Sci. U. S. A. 77(2): 919-23.

Kataoka, T., Miyata, T., Honjo, T. (1981): Repetitive sequences in class-switch recombination regions of immunoglobulin heavy chain genes. Cell 23(2): 357-68.

Kawakami, T., Takahashi, N., Honjo, T. (1980): Complete nucleotide sequence of mouse immunoglobulin mu gene and comparison with other immunoglobulin heavy chain genes. Nucleic. Acids. Res. 8(17): 3933-45.

Kawasaki, K., Minoshima, S., Nakato, E., Shibuya, K., Shintani, A., Asakawa, S., Sasaki, T., Klobeck, H.G., Combriato, G., Zachau, H.G., Shimizu, N. (2001): Evolutionary dynamics of the human immunoglobulin kappa locus and the germline repertoire of the Vkappa genes. Eur. J. Immunol. 31(4): 1017-28.

Kazazian, H.H. (2004): Mobile elements: drivers of genome evolution. Science 303(5664): 1626-32.

Kenter, A.L. (2003): Class-switch recombination: after the dawn of AID. Curr. Opin. Immunol. 15(2): 190-8.

Klein, U., Rajewsky, K., Kuppers, R. (1998): Human immunoglobulin (Ig)M+IgD+ peripheral blood B cells expressing the CD27 cell surface antigen carry somatically mutated variable region genes: CD27 as a general marker for somatically mutated (memory) B cells. J. Exp. Med. 188(9): 1679-89.

Kofler, R., Geley, S., Kofler, H., Helmberg, A. (1992): Mouse variable-region gene families: complexity, polymorphism and use in non-autoimmune responses. Immunol. Rev. 128: 5-21.

Kosak, S.T., Skok, J.A., Medina, K.L., Riblet, R., Le, B.e.a.u., Fisher, A.G., Singh, H. (2002): Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development. Science 296(5565): 158-62.

Krangel, M.S. (2003): Gene segment selection in V(D)J recombination: accessibility and beyond. Nat. Immunol. 4(7): 624-30.

Kurosawa, Y., Tonegawa, S. (1982): Organization, structure, and assembly of immunoglobulin heavy chain diversity DNA segments. J. Exp. Med. 155(1): 201-18.

Küppers, R., Dalla-Favera, R. (2001): Mechanisms of chromosomal translocations in B cell lymphomas. Oncogene 20(40): 5580-94.

Lawler, A.M., Lin, P.S., Gearhart, P.J. (1987): Adult B-cell repertoire is biased toward two heavy-chain variable-region genes that rearrange frequently in fetal pre-B cells. Proc. Natl. Acad. Sci. U. S. A. 84(8): 2454-8.

Lee, C.G., Kinoshita, K., Arudchandran, A., Cerritelli, S.M., Crouch, R.J., Honjo, T. (2001): Quantitative regulation of class switch recombination by switch region transcription. J. Exp. Med. 194(3): 365-74.

Lefranc, M.P., Giudicelli, V., Ginestoux, C., Bodmer, J., Müller, W., Bontrop, R., Lemaitre, M., Malik, A., Barbie, V., Chaume, D. (1999): IMGT, the international ImMunoGeneTics database. Nucleic. Acids. Res. 27(1): 209-12.

Lefranc, M.P., Giudicelli, V., Ginestoux, C., Bosc, N., Folch, G., Guiraudou, D., Jabado-Michaloud, J., Magris, S., Scaviner, D., Thouvenin, V., Combres, K., Girod, D., Jeanjean, S., Protat, C., Yousfi-Monod, M., Duprat, E., Kaas, Q., Pommie, C., Chaume, D., Lefranc, G. (2004): IMGT-ONTOLOGY for immunogenetics and immunoinformatics. In. Silico. Biol. 4(1): 17-29.

Lefranc, M.P., Giudicelli, V., Kaas, Q., Duprat, E., Jabado-Michaloud, J., Scaviner, D., Ginestoux, C., Clement, O., Chaume, D., Lefranc, G. (2005): IMGT, the international ImMunoGeneTics information system. Nucleic. Acids. Res. 33(Database issue): D593-7.

Lefranc, M.P., Lefranc, G. (2004): Immunoglobulin lambda (igl) genes of human and mouse. In: Molecular Biology of B cells. Honjo T, Alt FW, Neuberger M (ed); Elsevier Academic Press, London; 27-36.

Lefranc, M.P., Pommie, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L., Thouvenin-Contet, V., Lefranc, G. (2003): IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev. Comp. Immunol. 27(1): 55-77.

Lewis, S.M. (1994): P nucleotides, hairpin DNA and V(D)J joining: making the connection. Semin. Immunol. 6(3): 131-41.

Li, S.C., Rothman, P.B., Zhang, J., Chan, C., Hirsh, D., Alt, F.W. (1994): Expression of I mu-C gamma hybrid germline transcripts subsequent to immunoglobulin heavy chain class switching. Int. Immunol. 6(4): 491-7.

Litman, G.W., Cannon, J.P., Rast, J.P. (2005): New insights into alternative mechanisms of immune receptor diversification. Adv. Immunol. 87: 209-36.

Little, T.J., Hultmark, D., Read, A.F. (2005): Invertebrate immunity and the limits of mechanistic immunology. Nat. Immunol. 6(7): 651-4.

Livant, D., Blatt, C., Hood, L. (1986): One heavy chain variable region gene segment subfamily in the BALB/c mouse contains 500-1000 or more members. Cell 47(3): 461-70.

Lorenz, M., Jung, S., Radbruch, A. (1995): Switch transcripts in immunoglobulin class switching. Science 267(5205): 1825-8.

Lutzker, S., Alt, F.W. (1988): Structure and expression of germ line immunoglobulin gamma 2b transcripts. Mol. Cell. Biol. 8(4): 1849-52.

Mainville, C.A., Sheehan, K.M., Klaman, L.D., Giorgetti, C.A., Press, J.L., Brodeur, P.H. (1996): Deletional mapping of fifteen mouse VH gene families reveals a common organization for three Igh haplotypes. J. Immunol. 156(3): 1038-46.

Maizels, N. (2005): Immunoglobulin Gene Diversification. Annu. Rev. Genet. (elektrontische Publikation vor dem Druck)

Maki, R., Roeder, W., Traunecker, A., Sidman, C., Wabl, M., Raschke, W., Tonegawa, S. (1981): The role of DNA rearrangement and alternative RNA processing in the expression of immunoglobulin delta genes. Cell 24(2): 353-65.

Matsuda, F. (2004): Human immunoglobulin heavy chain locus. In: Molecular Biology of B cells. Honjo T, Alt FW, Neuberger M (ed); Elsevier Academic Press, London; 2-18.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D.U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S.,
Max, E.E. (1999): Immunoglobulins: Molecular Genetics. In: Fundamental Immunology. Paul (ed); Lippincott-Raven, Philadeplphia; 37-74.

Saxel, H., Scheer, M., Thiele, S., Wingender, E. (2003): TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic. Acids. Res. 31(1): 374-8.

Meek, K., Eversole, T., Capra, J.D. (1991): Conservation of the most JH proximal Ig VH gene segment (VHVI) throughout primate evolution. J. Immunol. 146(7): 2434-8.

Milstein, C., Brownlee, G.G., Harrison, T.M., Mathews, M.B. (1972): A possible precursor of immunoglobulin light chains. Nat. New. Biol. 239(91): 117-20.

Morea, V., Tramontano, A., Rustici, M., Chothia, C., Lesk, A.M. (1998): Conformations of the third hypervariable region in the VH domain of immunoglobulins. J. Mol. Biol. 275(2): 269-94.

Morgado, M.G., Cam, P., Gris-Liebe, C., Cazenave, P.A., Jouvin-Marche, E. (1989): Further evidence that BALB/c and C57BL/6 gamma 2a genes originate from two distinct isotypes. EMBO. J. 8(11): 3245-51.

Mostoslavsky, R., Alt, F.W., Rajewsky, K. (2004): The lingering enigma of the allelic exclusion mechanism. Cell 118(5): 539-44.

Mowatt, M., Dery, C., Dunnick, W. (1985): Unique sequences are interspersed among tandemly repeated elements in the murine gamma 1 switch segment. Nucleic. Acids. Res. 13(1): 225-37.

Müller, W., Nunes, M.P., Althaus, H.H.: DNAPLOT-Programm. http://www.dnaplot.de

Nemazee, D., Hogquist, K.A. (2003): Antigen receptor selection by editing or downregulation of V(D)J recombination. Curr. Opin. Immunol. 15(2): 182-9.

Neuberger, M.S., Di Noia J.M., Beale, R.C., Williams, G.T., Yang, Z., Rada, C. (2005): Somatic hypermutation at A.T pairs: polymerase error versus dUTP incorporation. Nat. Rev. Immunol. 5(2): 171-8.

Newell, N., Richards, J.E., Tucker, P.W., Blattner, F.R. (1980): J genes for heavy chain immunoglobulins of mouse. Science 209(4461): 1128-32.

Nikaido, T., Nakai, S., Honjo, T. (1981): Switch region of immunoglobulin Cmu gene is composed of simple tandem repetitive sequences. Nature 292(5826): 845-8.

Nossal, G.J. (1992): Cellular and molecular mechanisms of B lymphocyte tolerance. Adv. Immunol. 52: 283-331.

Oettinger, M.A. (2004): How to keep V(D)J recombination under control. Immunol. Rev. 200: 165-81.

Oettinger, M.A., Schatz, D.G., Gorka, C., Baltimore, D. (1990): RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. Science 248(4962): 1517-23.
Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., Yamanaka, I., Kiyosawa, H., Yagi, K., Tomaru, Y., Hasegawa, Y., Nogami, A., Schonbach, C., Gojobori, T., Baldarelli, R., Hill, D.P., Bult, C., Hume, D.A., Quackenbush, J., Schriml, L.M., Kanapin, A., Matsuda, H., Batalov, S., Beisel, K.W., Blake, J.A., Bradt, D., Brusic, V., Chothia, C., Corbani, L.E., Cousins, S., Dalla, E., Dragani, T.A., Fletcher, C.F., Forrest, A., Frazer, K.S., Gaasterland, T., Gariboldi, M., Gissi, C., Godzik, A., Gough, J., Grimmond, S., Gustincich, S., Hirokawa, N., Jackson, I.J., Jarvis, E.D., Kanai, A., Kawaji, H., Kawasawa, Y., Kedzierski, R.M., King, B.L., Konagaya, A., Kurochkin, I.V., Lee, Y., Lenhard, B., Lyons, P.A., Maglott, D.R., Maltais, L., Marchionni, L., McKenzie, L., Miki, H., Nagashima, T., Numata, K., Okido, T., Pavan, W.J., Pertea, G., Pesole, G., Petrovsky, N., Pillai, R., Pontius, J.U., Qi, D., Ramachandran, S., Ravasi, T., Reed, J.C., Reed, D.J., Reid, J., Ring, B.Z., Ringwald, M., Sandelin, A., Schneider, C., Semple, C.A., Setou, M., Shimada, K., Sultana, R., Takenaka, Y., Taylor, M.S., Teasdale, R.D., Tomita, M., Verardo, R., Wagner, L., Wahlestedt, C., Wang, Y., Watanabe, Y., Wells, C., Wilming, L.G., Wynshaw-Boris, A., Yanagisawa, M., Yang, I., Yang, L., Yuan, Z., Zavolan, M., Zhu, Y., Zimmer, A., Carninci, P., Hayatsu, N., Hirozane-Kishikawa, T., Konno, H., Nakamura, M., Sakazume, N., Sato, K., Shiraki, T., Waki, K., Kawai, J., Aizawa, K., Arakawa, T., Fukuda, S., Hara, A., Hashizume, W., Imotani, K., Ishii, Y., Itoh, M., Kagawa, I., Miyazaki, A., Sakai, K., Sasaki, D., Shibata, K., Shinagawa, A., Yasunishi, A., Yoshino, M., Waterston, R., Lander, E.S., Rogers, J., Birney, E., Hayashizaki, Y.;. (2002): Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420(6915): 563-73.

Ota, T., Nei, M. (1994): Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. Mol. Biol. Evol. 11(3): 469-82.

Pancer, Z., Amemiya, C.T., Ehrhardt, G.R., Ceitlin, J., Gartland, G.L., Cooper, M.D. (2004): Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. Nature 430(6996): 174-80.

Pancer, Z., Saha, N.R., Kasamatsu, J., Suzuki, T., Amemiya, C.T., Kasahara, M., Cooper, M.D. (2005): Variable lymphocyte receptors in hagfish. Proc. Natl. Acad. Sci. U. S. A. 102(26): 9224-9.

Petersen-Mahrt, S. (2005): DNA deamination in immunity. Immunol. Rev. 203: 80-97.

Radbruch, A., Burger, C., Klein, S., Müller, W. (1986): Control of immunoglobulin class switch recombination. Immunol. Rev. 89: 69-83.

Radcliffe, G., Lin, Y.C., Julius, M., Marcu, K.B., Stavnezer, J. (1990): Structure of germ line immunoglobulin alpha heavy-chain RNA and its location on polysomes. Mol. Cell. Biol. 10(1): 382-6.

Reth, M.G., Alt, F.W. (1984): Novel immunoglobulin heavy chains are produced from DJH gene segment rearrangements in lymphoid cells. Nature 312(5993): 418-23.

Reynaud, C.A., Aoufouchi, S., Faili, A., Weill, J.C. (2003): What role for AID: mutator, or assembler of the immunoglobulin mutasome? Nat. Immunol. 4(7): 631-8.

Riblet, R. (2004): Immunoglobulin Heavy Chain Genes of Mouse. In: Molecular Biology of B cells. Honjo T, Alt FW, Neuberger M (ed); Elsevier Academic Press, London; 19-26.

Rice, P., Longden, I., Bleasby, A. (2000): EMBOSS: the European Molecular Biology Open Software Suite. Trends. Genet. 16(6): 276-7.
Rogers, J., Choi, E., Souza, L., Carter, C., Word, C., Kuehl, M., Eisenberg, D., Wall, R. (1981): Gene segments encoding transmembrane carboxyl termini of immunoglobulin gamma chains. Cell 26(1 Pt 1): 19-27.

Rogers, J., Early, P., Carter, C., Calame, K., Bond, M., Hood, L., Wall, R. (1980): Two mRNAs with different 3' ends encode membrane-bound and secreted forms of immunoglobulin mu chain. Cell 20(2): 303-12.

Rohrbach, P., Broders, O., Toleikis, L., Dubel, S. (2003): Therapeutic antibodies and antibody fusion proteins. Biotechnol. Genet. Eng. Rev. 20: 137-63.

Roldan, E., Fuxa, M., Chong, W., Martinez, D., Novatchkova, M., Busslinger, M., Skok, J.A. (2005): Locus 'decontraction' and centromeric recruitment contribute to allelic exclusion of the immunoglobulin heavy-chain gene. Nat. Immunol. 6(1): 31-41.

Romo-Gonzalez, T., Vargas-Madrazo, E. (2005): Structural analysis of substitution patterns in alleles of human immunoglobulin VH genes. Mol. Immunol. 42(9): 1085-97.

Romo-Gonzalez, T., Vargas-Madrazo, E. (2005): Substitution patterns in alleles of immunoglobulin V genes in humans and mice. Mol. Immunol. Elektronische Veröffentlichung vor dem Druck.

Roth, D.B. (2003): Restraining the V(D)J recombinase. Nat. Rev. Immunol. 3(8): 656-66.

Sakano, H., Kurosawa, Y., Weigert, M., Tonegawa, S. (1981): Identification and nucleotide sequence of a diversity DNA segment (D) of immunoglobulin heavy-chain genes. Nature 290(5807): 562-5.

Sakano, H., Maki, R., Kurosawa, Y., Roeder, W., Tonegawa, S. (1980): Two types of somatic recombination are necessary for the generation of complete immunoglobulin heavy-chain genes. Nature 286(5774): 676-83.

Schatz, D.G. (2004): V(D)J recombination. Immunol. Rev. 200: 5-11.

Schatz, D.G., Oettinger, M.A., Baltimore, D. (1989): The V(D)J recombination activating gene, RAG-1. Cell 59(6): 1035-48.

Schebesta, M., Heavey, B., Busslinger, M. (2002): Transcriptional control of B-cell development. Curr. Opin. Immunol. 14(2): 216-23.

Schlissel, M.S. (2004): Regulation of activation and recombination of the murine Igkappa locus. Immunol. Rev. 200: 215-23.

Schmid, C.D., Praz, V., Delorenzi, M., Perier, R., Bucher, P. (2004): The Eukaryotic Promoter Database EPD: the impact of in silico primer extension. Nucleic. Acids. Res. 32(Database issue): D82-5.

Schmid, C.W. (1998): Does SINE evolution preclude Alu function? Nucleic. Acids. Res. 26(20): 4541-50.

Schroeder, H.W., Hillson, J.L., Perlmutter, R.M. (1990): Structure and evolution of mammalian VH families. Int. Immunol. 2(1): 41-50.

Schupp, I.W., Schlake, T., Kirschbaum, T., Zachau, H.G., Boehm, T. (1997): A yeast artificial chromosome contig spanning the mouse immunoglobulin kappa light chain locus. Immunogenetics 45(3): 180-7.

Schwartz, S., Elnitski, L., Li, M., Weirauch, M., Riemer, C., Smit, A., Green, E.D., Hardison, R.C., Miller, W.;. (2003): MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. Nucleic. Acids. Res. 31(13): 3518-24.

Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., Miller, W. (2003): Human-mouse alignments with BLASTZ. Genome. Res. 14(4): 786.

Shimizu, A., Takahashi, N., Yaoita, Y., Honjo, T. (1982): Organization of the constant-region gene family of the mouse immunoglobulin heavy chain. Cell 28(3): 499-506.

Sikorav, J.L., Auffray, C., Rougeon, F. (1980): Structure of the constant and 3' untranslated regions of the murine Balb/c gamma 2a heavy chain messenger RNA. Nucleic. Acids. Res. 8(14): 3143-55.

Silver, LM. (1995): Mouse Genetics. Oxford University Press. Internet-Version: http://www.informatics.jax.org/silver/

Sims, M.J., Krawinkel, U., Taussig, M.J. (1992): Characterization of germ-line genes of the VGAM3.8 VH gene family from BALB/c mice. J. Immunol. 149(5): 1642-8.

Sitnikova, T., Su, C. (1998): Coevolution of immunoglobulin heavy- and light-chain variable-region gene families. Mol. Biol. Evol. 15(6): 617-25.

Smit, A.F.A., Hubley, R., Green, P.: Repeatmasker. http://www.repeatmasker.org

Stavnezer, J. (2000): Molecular processes that regulate class switching. Curr. Top. Microbiol. Immunol. 245(2): 127-68.

Stavnezer, J., Amemiya, C.T. (2004): Evolution of isotype switching. Semin. Immunol. 16(4): 257-75.

Strausberg, R.L., Feingold, E.A., Grouse, L.H., Derge, J.G., Klausner, R.D., Collins, F.S., Wagner, L., Shenmen, C.M., Schuler, G.D., Altschul, S.F., Zeeberg, B., Buetow, K.H., Schaefer, C.F., Bhat, N.K., Hopkins, R.F., Jordan, H., Moore, T., Max, S.I., Wang, J., Hsieh, F., Diatchenko, L., Marusina, K., Farmer, A.A., Rubin, G.M., Hong, L., Stapleton, M., Soares, M.B., Bonaldo, M.F., Casavant, T.L., Scheetz, T.E., Brownstein, M.J., Usdin, T.B., Toshiyuki, S., Carninci, P., Prange, C., Raha, S.S., Loquellano, N.A., Peters, G.J., Abramson, R.D., Mullahy, S.J., Bosak, S.A., McEwan, P.J., McKernan, K.J., Malek, J.A., Gunaratne, P.H., Richards, S., Worley, K.C., Hale, S., Garcia, A.M., Gay, L.J., Hulyk, S.W., Villalon, D.K., Muzny, D.M., Sodergren, E.J., Lu, X., Gibbs, R.A., Fahey, J., Helton, E., Ketteman, M., Madan, A., Rodrigues, S., Sanchez, A., Whiting, M., Madan, A., Young, A.C., Shevchenko, Y., Bouffard, G.G., Blakesley, R.W., Touchman, J.W., Green, E.D., Dickson, M.C., Rodriguez, A.C., Grimwood, J., Schmutz, J., Myers, R.M., Butterfield, Y.S., Krzywinski, M.I., Skalska, U., Smailus, D.E., Schnerch, A., Schein, J.E., Jones, S.J., Marra, M.A.;. (2002): Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. Proc.

Natl. Acad. Sci. U. S. A. 99(26): 16899-903.

Su, I.H., Tarakhovsky, A. (2005): Epigenetic control of B cell differentiation. Semin. Immunol. 17(2): 167-72.

Su, I.H., Tarakhovsky, A. (2005): Epigenetic control of B cell differentiation. Semin. Immunol. 17(2): 167-72.

Swanson, P.C. (2004): The bounty of RAGs: recombination signal complexes and reaction outcomes. Immunol. Rev. 200: 90-114.

Szurek, P., Petrini, J., Dunnick, W. (1985): Complete nucleotide sequence of the murine gamma 3 switch region and analysis of switch recombination sites in two gamma 3-expressing hybridomas. J. Immunol. 135(1): 620-6.

Takahashi, N., Kataoka, T., Honjo, T. (1980): Nucleotide sequences of class-switch recombination region of the mouse immunoglobulin gamma 2b-chain gene. Gene 11(1-2): 117-27.

Thiebe, R., Schable, K.F., Bensch, A., Brensing-Kuppers, J., Heim, V., Kirschbaum, T., Mitlohner, H., Ohnrich, M., Pourrajabi, S., Roschenthaler, F., Schwendinger, J., Wichelhaus, D., Zocher, I., Zachau, H.G. (1999): The variable genes and gene families of the mouse immunoglobulin kappa locus. Eur. J. Immunol. 29(7): 2072-81.

Tompa, M., Li, N., Bailey, T.L., Church, G.M., De, M.o.o.r., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., Makeev, V.J., Mironov, A.A., Noble, W.S., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., van, H.e.l.d.e.n., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., Zhu, Z. (2005): Assessing computational tools for the discovery of transcription factor binding sites. Nat. Biotechnol. 23(1): 137-44.

Tonegawa, S. (1983): Somatic generation of antibody diversity. Nature 302(5909): 575-81.

Trepicchio, W.Jr., Barrett, K.J. (1985): The Igh-V locus of MRL mice: restriction fragment length polymorphism in eleven strains of mice as determined with VH and D gene probes. J. Immunol. 134(4):2734-9.

Tucker, P.W., Liu, C.P., Mushinski, J.F., Blattner, F.R. (1980): Mouse immunoglobulin D: messenger RNA and genomic DNA sequences. Science 209(4463): 1353-60.

Tucker, P.W., Slightom, J.L., Blattner, F.R. (1981): Mouse IgA heavy chain gene sequence: implications for evolution of immunoglobulin hinge axons. Proc. Natl. Acad. Sci. U. S. A. 78(12): 7684-8.

Tutter, A., Riblet, R. (1988): Duplications and deletions of Vh genes in inbred strains of mice. Immunogenetics 28(2): 125-35.

Van Snick, J.L., Masson, P.L. (1979): Age-dependent production of IgA and IgM autoantibodies against IgG2a in a colony of 129/Sv mice. J. Exp. Med. 149(6): 1519-30.

Wang, C.L., Wabl, M. (2004): DNA acrobats of the Ig class switch. J. Immunol. 172(10): 5815-21.

Ward, S.B., Morrison, S.L. (1992): Sequence of the gamma 2b membrane 3' untranslated region: polya site determination and comparison to the gamma 2a membrane 3' untranslated region. Mol. Immunol. 29(2): 279-85.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S.E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M.R., Brown, D.G., Brown, S.D., Bult, C., Burton, J., Butler, J., Campbell, R.D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A.T., Church, D.M., Clamp, M., Clee, C., Collins, F.S., Cook, L.L., Copley, R.R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K.D., Deri, J., Dermitzakis, E.T., Dewey, C., Dickens, N.J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D.M., Eddy, S.R., Elnitski, L., Emes, R.D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G.A., Flicek, P., Foley, K., Frankel, W.N., Fulton, L.A., Fulton, R.S., Furey, T.S., Gage, D., Gibbs, R.A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T.A., Green, E.D., Gregory, S., Guigo, R., Guyer, M., Hardison, R.C., Haussler, D., Hayashizaki, Y., Hillier, L.W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D.B., Johnson, L.S., Jones, M., Jones, T.A., Joy, A., Kamal, M., Karlsson, E.K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W.J., Kirby, A., Kolbe, D.L., Korf, I., Kucherlapati, R.S., Kulbokas, E.J., Kulp, D., Landers, T., Leger, J.P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D.R., Mardis, E.R., Matthews, L., Mauceli, E., Mayer, J.H., McCarthy, M., McCombie, W.R., McLaren, S., McLay, K., McPherson, J.D., Meldrim, J., Meredith, B., Mesirov, J.P., Miller, W., Miner, T.L., Mongin, E., Montgomery, K.T., Morgan, M., Mott, R., Mullikin, J.C., Muzny, D.M., Nash, W.E., Nelson, J.O., Nhan, M.N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M.J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K.H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C.S., Poliakov, A., Ponce, T.C., Ponting, C.P., Potter, S., Quail, M., Reymond, A., Roe, B.A., Roskin, K.M., Rubin, E.M., Rust, A.G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M.S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J.B., Slater, G., Smit, A., Smith, D.R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J.P., Von, N.i.e.d.e.r.h.a.u.s.e.r.n., Wade, C.M., Wall, M., Weber, R.J., Weiss, R.B., Wendl, M.C., West, A.P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R.K., Winter, E., Worley, K.C., Wyman, D., Yang, S., Yang, S.P., Zdobnov, E.M., Zody, M.C., Lander, E.S.;. (2002): Initial sequencing and comparative analysis of the mouse genome. Nature 420(6915): 520-62.

Watson, F.L., Puttmann-Holgado, R., Thomas, F., Lamar, D.L., Hughes, M., Kondo, M., Rebel, V.I., Schmucker, D. (2005): Extensive Diversity of Ig-Superfamily Proteins in the Immune System of Insects. Science. Elektronische Veröffentlichung vor dem Druck.

Whitcomb, E.A., Haines, B.B., Parmelee, A.P., Pearlman, A.M., Brodeur, P.H. (1999): Germline structure and differential utilization of Igha and Ighb VH10 genes. J. Immunol. 162(3): 1541-50.

Williams, G.S., Martinez, A., Montalbano, A., Tang, A., Mauhar, A., Ogwaro, K.M., Merz, D., Chevillard, C., Riblet, R., Feeney, A.J. (2001): Unequal VH gene rearrangement frequency within the large VH7183 gene family is not due to recombination signal sequence variation, and mapping of the genes shows a bias of rearrangement based on chromosomal location. J. Immunol. 167(1): 257-63.

Word, C.J., Mushinski, J.F., Tucker, P.W. (1983): The murine immunoglobulin alpha gene expresses multiple transcripts from a unique membrane exon. EMBO. J. 2(6): 887-98.

Wu, T.T., Kabat, E.A. (1970): An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. J. Exp. Med. 132(2): 211-50.

Wu, T.T., Reid-Miller, M., Perry, H.M., Kabat, E.A. (1984): Long identical repeats in the mouse gamma 2b switch region and their implications for the mechanism of class switching. EMBO. J. 3(9): 2033-40.

Yamawaki-Kataoka, Y., Nakai, S., Miyata, T., Honjo, T. (1982): Nucleotide sequences of gene segments encoding membrane domains of immunoglobulin gamma chains. Proc. Natl. Acad. Sci. U. S. A. 79(8): 2623-7.

Ye, J. (2004): The immunoglobulin IGHD gene locus in C57BL/6 mice. Immunogenetics 56(6): 399-404.

Yu, K., Lieber, M.R. (2003): Nucleic acid structures and enzymes in the immunoglobulin class switch recombination mechanism. DNA. Repair. (Amst). 2(11): 1163-74.

Zachau, H.G. (2004): Immunoglobulin k genes of human and mouse. In: Molecular Biology of B cells. Honjo T, Alt FW, Neuberger M (ed); Elsevier Academic Press, London; 27-36.

Zhang, S.M., Adema, C.M., Kepler, T.B., Loker, E.S. (2004): Diversification of Ig superfamily genes in an invertebrate. Science 305(5681): 251-4.

Danksagung

Ich möchte mich bei allen Menschen bedanken, die mich bei der Anfertigung meiner Dissertation unterstützt haben.

Zunächst bedanke ich mich bei Professor Jürgen Wehland und Professor Stefan Dübel für die Übernahme der Begutachtung dieser Arbeit.

Ein ganz besonderer Dank gilt Werner Müller für die hervorragende Betreuung und für das große Vertrauen in meine Lernfähigkeit.

Bei der Abteilung Genomanalyse von Helmut Blöcker und insbesondere bei Maren Scharfe möchte ich mich für die gute Kooperation und die Bereitstellung der genomischen Sequenzen bedanken.

In Bezug auf die VBASE2-Datenbank möchte ich Richard Münch, Miguel Nunes und Andreas Kahari ganz herzlich für die ausgesprochen konstruktive Zusammenarbeit und Unterstützung danken.

Für die Betreuung meines Marie-Curie-Aufenthalts am Europäischen Bioinformatik-Institut bedanke ich mich bei Maria Jesus-Martin, Claire O'Donovan und Rolf Apweiler.

Alle Mitglieder der Abteilung Experimentelle Immunologie haben ein herzliches Dankeschön für die Hilfsbereitschaft und freundliche Arbeitsatmosphäre verdient. Ganz besonders erwähnen möchte ich Rolf Hühne, der mir beim Einstieg in die Welt der Bioinformatik sehr geholfen hat, und Martin Hafner, der jederzeit für interessante Diskussionen zur Verfügung stand und mich auch stets mit der dazugehörigen Literatur versorgte.

Ausdrücklich erwähnen möchte ich auf dieser Seite auch unsere Tagesmutter Hannah Kartheuser, die durch ihre jahrelange professionelle und zuverlässige Arbeit meine Dissertation überhaupt erst ermöglicht hat.

Nicht zuletzt gilt mein Dank meinem Mann Heiko Rabba und meinen Kindern Lisa und Tim, die wirklich sehr viel Geduld mit mir und meiner Arbeit hatten.